UHASSELT

KNOWLEDGE IN ACTION

Faculty of Business Economics

Master of Management

Master's thesis

Organizational factors that influence fairness in algorithmic decision-support

Deborah Adelakun

Thesis presented in fulfillment of the requirements for the degree of Master of Management, specialization Business Process Management

SUPERVISOR :

Prof. dr. Koenraad VANHOOF

MENTOR :

Mevrouw Elisavet KOUTSOVITI-KOUMERI

2023
2024

UHASSELT

KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT

KNOWLEDGE IN ACTION

Faculty of Business Economics

Master of Management

Master's thesis

Organizational factors that influence fairness in algorithmic decision-support

Deborah Adelakun

Thesis presented in fulfillment of the requirements for the degree of Master of Management, specialization Business Process Management

SUPERVISOR :

Prof. dr. Koenraad VANHOOF

MENTOR :

Mevrouw Elisavet KOUTSOVITI-KOUMERI

Disclaimer

I hereby declare that this thesis is my original work and has not been submitted to any institution for assessment purposes. All sources used have been acknowledged and cited in the reference section.

ACKNOWLEDGEMENT

My sincere appreciation goes to Lisa Koutsoviti, who collaborated with Prof. Vanhoof Koen. Your invaluable guidance, generosity with your time, contributions, and availability have been instrumental. Your expertise, patience, insight, and attention to detail are deeply appreciated. You truly know your onions and you deserve more than flowers.

I would like to extend my heartfelt thanks to my husband Gaius (Ololufe) for your unconditional support, for always believing in me, and for taking on other responsibilities so I could complete this work. My sweet baby, Akachi, your background presence during some of my meetings added a beautiful spice to my life 😊.

Special thanks to my sister Mercy for always being there when I needed help, and to Dickson for your technical expertise and insights. To my brother Oluwadara, thank you for always checking up on me.

To all my friends and family whose counsel and support have guided my journey, I say thank you. You are heard, loved, and seen: Dr. Adesoji Oluyomi mi owon, Egbon Opeyemi David, John Ayanfe, and many more.

ABSTRACT

Algorithmic Decision Support (ADS) is rapidly becoming an integral part of daily life, utilized by individuals, organizations in both public and private sectors, governments, and businesses. Despite the growing adoption of ADS, there is an increasing concern about issues such as bias, accountability, and fairness. Not enough research focuses on organizational responsibilities in designing and adopting fair ADS systems. While this is a nascent area of research, much remains to be explored. This study systematically reviewed 28 articles, including both empirical research and literature reviews, to highlight the organizational factors influencing fairness during both the design and implementation processes. Recommendations from the literature were synthesized to suggest actions organizations can take to improve the fairness of ADS systems. The review identified key factors that can enhance fairness and outlined the responsibilities of organizations in implementing tools to promote fairness throughout the development and deployment stages. The study concluded that although addressing fairness in ADS is challenging due to its sociotechnical nature requiring a multifaceted approach, it is possible to influence fairness by incorporating it into the organizational goals from the outset.

Keywords:
Algorithmic Decision-Support; Algorithm Fairness; Organisation.

LIST OF TABLES

LIST OF FIGURES

TABLE OF CONTENT

1.0. INTRODUCTION

With the rise in the use of Artificial Intelligence (AI) and the increasing reliance on AI to assist in human decision-making (Marcinkowski *et al.,* 2020; Araujo *et al.,* 2020; Schumann *et al.,* 2020; Vlasceanu *et al.,* 2022; Kirkpatrick, 2016) which is becoming prevalent in modern work environments, it is crucial to understand the systems that support these decisions. This need has led to calls for explainable systems, as an explainable ADS is perceived as more fair (Munch *et al.,* 2024; Koutsouris, n.d.). Algorithmic Decision-Support (ADS) refers to the use of algorithms to process data and support or automate decision-making processes in various domains, leveraging computational capabilities to automate and standardize decisions. ADS plays a crucial role in strategic decision-making by providing timely and accurate assessments of key organizational aspects (Marabelli *et al.,* 2021; Bader & Kaiser, 2019; Lee, 2019). However, the use of ADS raises concerns about ethical and discussion on its fairness (Marcinkowski *et al.,* 2020; Woodruff *et al.,* 2018).

Fair Algorithmic Decision-Support (ADS) involves the design, implementation, and use of algorithms in ways that ensure decisions do not produce unjust, discriminatory, or disparate outcomes (Lee, 2018). Fair ADS aims to prevent algorithms from reinforcing biases and to foster equity, accountability, and transparency in decision-making processes (Starke *et al.,* 2022). To mitigate errors and maximize the benefits of ADS without perpetuating biases and injustices, fair algorithm design and implementation are necessary. Achieving fairness in ADS is complex due to biases in data, the intricacies of algorithm design, the complexity of fairness metrics, the integration of technical and ethical considerations, and the need for transparency and contextual sensitivity (Wang *et al.,* 2022; Koutsouris, n.d.). Organizations are responsible for the design, application, outcomes, impact, and effects of their systems (Hermann, 2022). Therefore, they must prioritize fair algorithmic decision support (ADS) to comply with legal standards, fulfill ethical responsibilities, build trust, and enhance decision quality, ultimately contributing to a fairer society. To fully leverage the benefits of ADS, it is crucial to identify the risks it poses and ensure responsible design and use of the system across various contexts (Adensamer *et al.,* 2021).

While various research areas explore the issue of accountability and user perception of algorithms, there is limited research on organizational responsibilities in developing and implementing fair ADS systems. Some organizations remain unaware of the effects of these systems until they are disseminated (Veale *et al.,* 2018), but the outcomes are still their responsibility. This underscores the need for meticulous inspection of compliance with standards to develop fair ADS. Several studies have approached fairness from a socio-technical perspective and developed tools to address this issue. For instance, research by Lee and Singh (2021), Ferrara et al. (2023), Metcalf et al. (2021), and Rana et al. (2023) suggest using assessment tools to tackle fairness challenges. However, there is a lack of research on how organizational (managerial) practices influence fairness. Understanding the role of management is crucial for supporting practitioners and industry experts in developing, building, and deploying fair ADS. When organizations take their responsibilities seriously and address unfairness, they promote justice and equity in their operations. An unfair ADS could pose challenges for organizations, but designing and implementing a fair ADS reduces legal risks and ensures compliance with anti-discrimination laws, protecting organizations from potential legal

repercussions. Moreover, it builds trust among stakeholders, such as customers, employees, and the broader public, enhancing the organization's reputation and credibility. Improving fairness in ADS systems by identifying relevant factors that promote fairness will also ai d organizations in contributing positively to social equity, reducing biases and disparities in high-impact areas, as organizations often face socio-technical problems with their design(Selbst *et al.,* 2019).

To identify factors that could help organizations improve ADS systems, a systematic literature review was conducted on twenty-eight selected articles relevant to the research objectives. The articles were systematically reviewed and synthesized to summarize existing factors influencing fairness in algorithmic decision support, with a focus on organizational factors. By understanding the influence of organizational factors on algorithmic decision support fairness, stakeholders can focus on designing, improving, and advancing effective strategies that impact and support fairness. The study also aims to identify and describe available knowledge, existing research gaps, and limitations of prior literature, providing reference ideas for future research informed by the findings of the research. The following research questions guided the review:

1. What are the organizational factors that influence fairness in algorithmic decision support?
2. What are the research gaps and limitations of the prior literature, and what future research opportunities can advance fairness in algorithmic decision support systems?

Addressing organizational factors influencing fairness is vital for creating fairer systems. By synthesizing the identified factors, organizations can mitigate biases and promote fairness in decision-making processes. Achieving fairness in algorithmic decision support (ADS) involves navigating various stages of the design process, each influenced by a variety of factors. From initial development to final deployment, the complex interplay of socio-contextual and socio-technical elements presents both challenges and opportunities. This study highlights factors that influence fairness in designing and implementing ADS. The factors were classified from two perspectives: theme-based and implementation-based. The theme-based classification includes:

- **Governance**: Policies and regulations that guide ADS development and use.
- **Social Responsibility**: Commitment to ethical principles and social impact.
- **Technical**: Ensuring algorithmic transparency, explainability, and accuracy in areas such as data management and human judgment during design integration and of ADS system.
- **Training & Development**: Educating stakeholders about ADS, expected ethical approach and its implications – Transparency, explainability, accountability.

The implementation-based classification was grouped into three categories:

1. **Organizational Related**: Internal policies and culture affecting ADS.
2. **Leadership Related**: Roles of leaders in promoting and overseeing fairness.
3. **Task Related**: Specific roles and responsibilities of practitioners involved in development and deployment.

The major factors identified include transparency, collaboration, data management, human judgment (Human-in-the-Loop), explainability, feature selection, communication, and accountability. Organizations can consider these factors throughout the development and implementation stages to achieve fairer outcomes. Identifying the appropriate stages and responsible individuals for implementing these factors is also essential. Organizations need to create support avenues for practitioners to ensure fair design outcomes, and leaders in each sector should integrate fairness metrics and guidelines to develop efficient, effective, and fair ADS. By focusing on these factors, organizations can:

- **Mitigate Biases**: Actively work to reduce biases in decision-support systems,
- **Promote Equity**: Ensure decisions are fair and non-discriminatory,
- **Build Trust**: Enhance reputation and credibility by demonstrating a commitment to fairness and,
- **Comply with Legal Standards**: Avoid legal repercussions by adhering to anti-discrimination laws.

By highlighting how organizational actions influence fairness, this review contributes to existing literature by identifying steps that can be taken to improve ADS systems during both the development and deployment phases. This includes actions by both the organizations designing the systems and those adopting and integrating them.

The review emphasizes the importance of organizational support for practitioners to ensure the creation of fair systems and addresses the challenges they face in this process. Ultimately, improving fairness in ADS systems contributes to social equity, reduces biases and disparities, and supports the broader goal of creating a just and equitable society. The insights from this systematic review provide valuable guidance for organizations seeking to enhance the fairness of their ADS systems and fulfill their ethical responsibilities.

The literature review provides a comprehensive overview of articles on fairness, accountability, and organizational responsibilities in algorithmic decision support systems (ADS). It is organized into five chapters: Chapter 1 introduces the topic, followed by Chapter 2, which outlines the methodology in detail - a systematic review of 28 different articles, including twelve empirical studies, fifteen literature reviews, and one study with a mixed methodology. Chapter 3 provides background to the study, including a review of existing literature on fairness in ADS. Chapter 4 discusses the findings from the literature on organizational factors influencing fairness. The final chapter, Chapter 5, presents the conclusion. Each chapter builds on the previous one to offer a thorough understanding of the subject.

2.0. METHODOLOGY

This study utilizes a systematic review methodology inspired by Wieringa's (2020) innovative approach, which is documented in the search strategy. A systematic review is structured around a defined research question and seeks to provide a comprehensive, transparent overview of existing knowledge. This review specifically sheds light on the organizational factors influencing fairness in algorithmic decision support by identifying and evaluating relevant literature on the topic. The study comprises three main steps: planning, conducting the review, and documenting the review, each with its own set of sub-steps as shown in Figure 1. (Research Design).

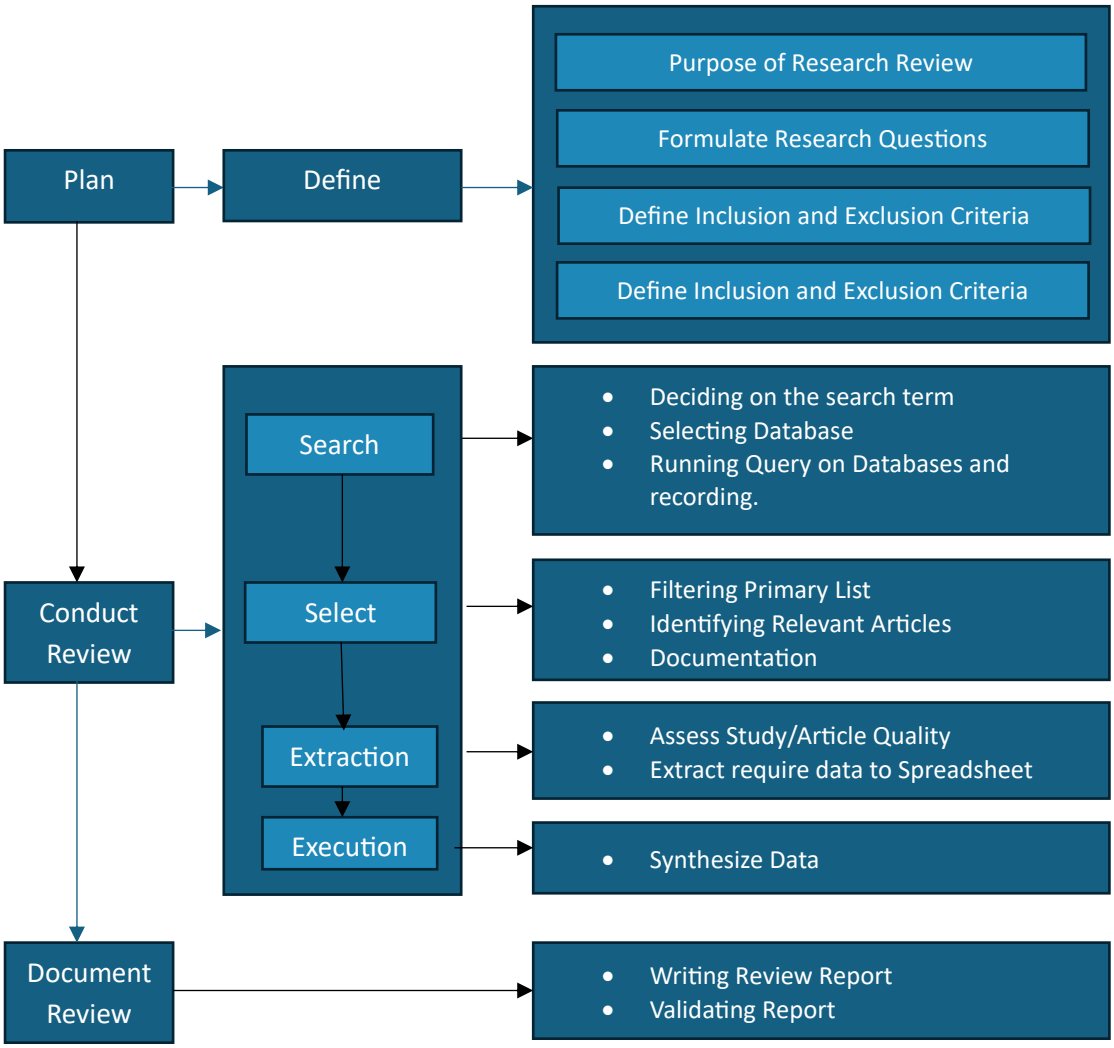

Figure 1: Research Design

2.1. Research Question

To maintain focus and ensure clarity, the research questions were developed based on the PICOC as shown in Table (Population, Intervention, Comparison, Outcome, Context) framework, guiding the formulation of the research questions.

RQ. 1. What are the organisational factors that influence fairness in algorithmic decision support?
RQ. 2. What are the research gaps and limitations of the prior literature and what future research opportunities can be derived to advance fairness in algorithmic decisions support systems?

Table 1: Summary of PICOC Framework

Research Question Formulated using **PICOC** Framework	
Population	Articles in which factors influencing fairness in Algorithmic Decision-Support concerning Management and organization were discussed.
Intervention	Highlighting a complete picture of organization related factors, practices, policies or strategies that may contribute to fairness in the use of Algorithmic decision support.
Comparison	Compare organizational factors that influence fairness in Algorithmic decision-support in different sectors.
Outcome	Comprehensive documentation of Algorithmic decision-support influencing fairness factors relating to organizational policy, structure, and practices and its impact /effect.
Context	Organizational context, including structure, culture, policies, etc. that may influence fairness in how algorithm decision-support is designed, implemented, or managed within the organization. Also, future research direction and literature gaps.

2.2. Study Selection Criteria

To ensure the review's foundation on quality evidence, inclusion criteria, and exclusion criteria were established. Only studies meeting these predefined criteria were considered for review. The Inclusion and exclusion criteria are shown in Table 2.

Table 2: Study Selection Criteria

Inclusion Criteria	**Exclusion Criteria**
The study has to discuss fairness in algorithmic decision support.	Studies/research that discuss algorithms without the decision part.
The study has to discuss factors that influence algorithmic decision support.	Studies/research that includes fairness but not in relation to fairness in algorithm decision support.
All industries	Studies that discuss Algorithmic decision-support but does not discuss fairness aspect.

The study/articles have to be written in English.	
The full study/article is accessible.	

2.3. Search strategy

The search strategy involves three sequential steps: selecting search terms, choosing databases, and executing search queries in the chosen databases, adjusting the search string as needed for each database's requirements.

2.31. Search Terms

Inspired by the work of Wieringa (2020), this research employed an exploratory query based on three pre-identified articles (Veale *et al.,* 2018; Madaio *et al.,* 2022; Holstein *et al.,* 2019). To accommodate the diversity of studies, related terms and antonyms were identified. Based on the keywords from the pre-selected papers, a search string was constructed to collect new articles. Initially, the search strings returned several generic articles that weren't related to the study objectives. Therefore, the search strings were adjusted base and refined to develop a final version that successfully retrieved relevant data.

2.32. Selecting Databases and Search Process

Four databases were selected for this study: ACM Digital Library, ScienceDirect, Scopus, and Springer. The search strings were run in each of these databases. The search was limited to the timeframe of 2000 to 2024 and included only articles in English. To ensure comprehensive search results, Boolean operators (AND, OR). For approximate phrases (wildcards and quotation marks) were employed. The search results were then exported to Endnote for further review and analysis, selected studies were exported to Excel.

2.33. Study Selection Strategy

The database search returned a total of 4,512 studies. The results were filtered by reviewing the titles and abstracts for relevance. The study selection process, including the number of articles removed based on selection criteria and the number of retained studies, is documented in the PRISMA chart shown in Figure 2.

The search string used for the databases was ("Algorithm* Accountability" AND "Algorithm* Bias" AND "Fairness" AND "*Decision Support" OR "Algorithm* Decision making"), except for Scopus, which does not accept wildcards. The modified search string used for Scopus was ("Algorithm Bias" AND "Fairness" AND "Algorithm Decision Support" OR "Algorithmic Decision making" OR "Algorithm Decision making").

Five (5) studies could not be retrieved. A secondary screening was conducted on seventy-two (72) studies, of which thirty-eight (38) were removed for lack of relevance, and two (2) were removed for wrong outcomes. Altogether, twenty-eight (28) papers were included in the study.

The selected studies were read to identify their methodologies and understand the different angles from which the topics have been explored. The objectives, results, and recommendations from the studies were analyzed to gain a comprehensive understanding of the ideas presented in the research.

2.34. Research Process steps

The studies were searched to ensure they included the keywords relevant to the research, particularly "*Fair**", "*Organization", "Decision-support*" and "*Algorithm**". The following protocols were followed for the research:

1. Search string input into the data bases one after the other.
2. The full record was downloaded.
3. The record was imported into Endnote for organization and Management.
4. Title and abstract reviewed according to the inclusion and exclusion criteria.
5. Download and review the full text to assess its relevancy to the study.
6. Record the studies that meet the inclusion criteria.
7. Document the findings of the selected studies in Excel Spreadsheet for further analysis.
8. Extract the organisational factors that can influence fairness from each of the articles
9. Extract the recommendations from the studies
10. Create a synthesis of the recommendations into a cohesive summary.

The outcomes and results of the above procedures are discussed in the next chapter.

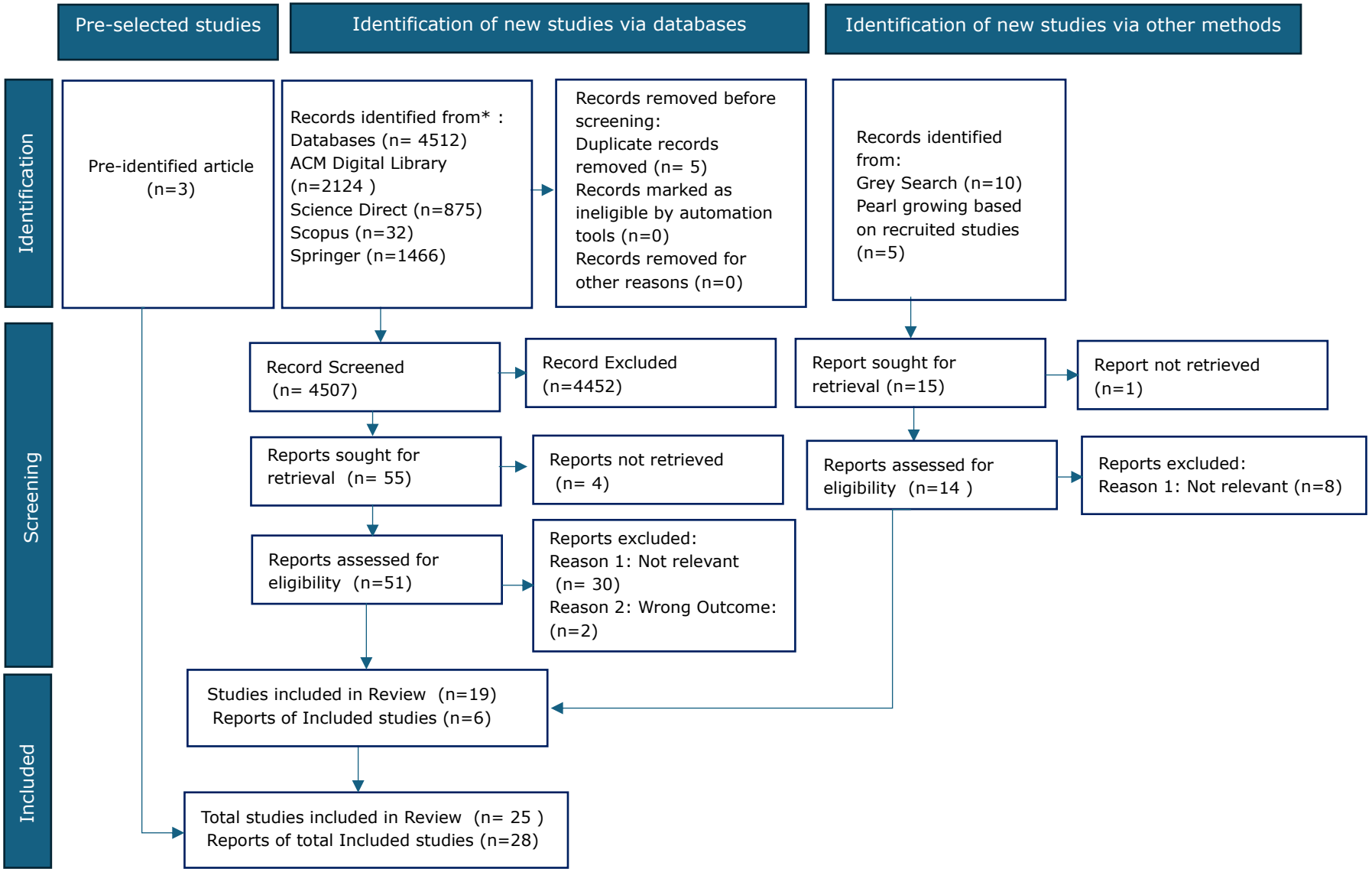

Figure 2: PRISMA

3.0 LITERATURE REVIEW

There is increasing research into Algorithmic decision-support, to address the issue of responsibility and accountability to increase fairness in Algorithmic decision-support. However, according to the information available the organisational factors that influence fairness have not been systematically examined; This section gives insight into the background and motivation of the study based on related research in context.

3.1 Introduction

Standardization of everyday business decisions, remote control, and automated decision-support are all considered aspects of algorithmic decision-support (Mohlmann and Zalmanson 2017). Algorithmic decision-support is primarily motivated by the aim to streamline operations, cut costs, reduce time constraints, mitigate risks, enhance productivity, and bolster confidence in decision-making processes. (Suen *et al.* 2019; McDonald *et al.* 2017; McColl and Michelotti 2019; Woods *et al.* 2020). For example, when algorithms, rather than humans, determine decisions, it has considerable ramifications for individuals and society in the context of optimizing organizations (Chalfin *et al.* 2016; Lee 2018; Lindebaum *et al.,* 2019). The shift toward algorithmic decision-making facilitates the automatic evaluation of numerous applications, enabling Human Resources to identify hidden talent within organizations more effectively. (Carey and Smith 2016; Silverman and Waller 2015; Savage and Bales 2017).

Apart from these profit-driven objectives, companies employ algorithmic decision-support to mitigate human biases (personal opinions and prejudices), enhancing the impartiality, uniformity, and equity of HR development and hiring practices (Langer *et al.* 2019; Florentine, 2016; Raghavan *et al.* 2020). The application of algorithmic decision support is expanding annually, influencing choices that profoundly affect the lives of individuals in a variety of fields, including human resources (Dreyer & Schulz, 2019), credit and welfare access (O'Neil, 2016), sentencing (Christin, 2017), and policing (Bennett Moses & Chan, 2018; Ferguson, 2017). This technology is generally regarded as having an edge over humans in decision-making because it is thought to be especially capable of using additional processing capacity and utilizing additional data to make conclusions that are more impartial and value-free than those made by humans (Christin, 2017; Gillespie, 2016).

However, it is possible for bias and unfairness to arise when relying solely on algorithmic decision-support (e.g., (Lee 2018; Lindebaum *et al.* 2019; Simbeck 2019)). Prejudice is commonly understood to be the unfair treatment of different groups according to age, gender, or ethnicity compared to more qualitative distinctions like Independent effectiveness (Arrow 1973). Input data that is unrepresentative (Suresh and Guttag, 2019), prejudiced (Barocas and Selbst 2016), or inaccurate (Kim 2016) are training grounds for algorithms that lead to discrimination or biased outcomes. In light of this, algorithms that rely on biased input (or training) data are susceptible to making or reproducing biased judgments (Chander 2016).

ADS systems may function fairly for certain tasks but poorly for others if they are haphazardly chosen, constructed, and described (Veale and Binns, 2017). Owing to unprecedented levels of data availability and processing capacity, both public and private organizations are using algorithms more frequently to make critical decisions like selecting qualified candidates, assigning patients to therapy, or forecasting criminal activity (AlgorithmWatch, 2019). For example, ADM systems have wrongly decreased residents' disability benefits, arbitrarily denied them access to food assistance programs, or unjustly charged them of being fraudsters (Richardson *et al.,* 2019).

ADS biases can arise from a variety of sources and are frequently inadvertent. These can happen during the process of gathering and analyzing input data as well as when choosing and defining the algorithm (Veale and Binns, 2017). First, ADS systems utilizing previous input data as training data are probably going to reinforce social biases already in place or even create new ones, frequently with negative effects on minority populations (Eubanks, 2018; Lepri *et al.,* 2018). Certain groups may be misrepresented due to inadequate or faulty data about individual or group characteristics (Köchling and Wehner, 2020). Furthermore, algorithms might be biased if they are haphazardly chosen, created, and defined. This is because some ADM systems might function well in certain tasks but inadequately in others (Veale and Binns, 2017).

The increasing adoption of algorithms in managerial and organizational decision-making is primarily driven by the potential to enable effective, optimal, and data-driven choices. However, regardless of the quality of decision outcomes, the delegation of judgment to algorithms rather than humans may influence how decisions are perceived by individuals (Sundar and Nass, 2001). These perceptions can impact attitudes and trust levels in algorithmic decision-making, which are crucial for thriving communities, businesses, and societies. For example, Skarlicki and Folger's 1997 research suggests that employees who perceive decisions made by managers or organizations as unfair may harbor resentment, exhibit anger, engage in retaliatory behavior, or even act against the organization.

According to published research, managerial strategies including recruiting experienced programmers, training them for regular maintenance, and validating datasets are crucial for reducing bias in algorithmic decision (Noriega, M., 2020). In order to oversee and manage algorithmic fairness, companies would need to create updated, modernized internal structures that are just and moral, as well as corporate strategies (Johnson, K. N., 2019). Additionally, there is a great need for organizational strategies aimed at increasing awareness of ethical and responsible AI as organisations are responsible for the result of their products. These strategies should prioritize workforce diversity within the organization. Policies that support fairness and incorporate cultural diversity into data can also help to counteract algorithmic bias (Lee, N. T., 2018).

AI fairness transcends mere technical considerations and encompasses socio-technical dimensions that can be complemented by effective managerial techniques. In a socio-technical framework, algorithms serve as standards and heuristic guides to benefit users and advance society. Rather than viewing algorithms as flawless entities or abstract concepts, a socio-technical analysis delves into the social and human decisions underlying these technological advancements (Shin & Choi, 2014). Recognizing algorithms as part of a socio-technical ecosystem facilitates the transition

towards human-centered and sustainable algorithmic usage in society (MacKenzie, 2014). AI Fairness is a socio-technical issue rather than just a technical one and could be coupled with efficient managerial techniques. In this literature review, we will further discuss the factors influencing fairness in algorithmic decision support, and guidelines for ethical design and implementation.

3.2 Factors Influencing Fairness in Algorithmic Decision-Support

Algorithmic biases may result from preexisting biases in the real-world systems that the data measures. The phrase "bias in, bias out" refers to this particular source of bias, which is the one that is most frequently mentioned in public conversations regarding algorithmic bias (Rambachan & Roth, 2020; Mayson, 2018; Courtland, 2018). The incoming data usually contains biases from the measured world, and the generated models that reflect those biases will also typically have those biases. For instance, our historical data will demonstrate that people who are part of identified minorities have a lower probability of achieving success even in situations where other variables are identical if systemic racism at a university affects student success. A predictive model created using a typical learning algorithm will discover that these students have a reduced chance of succeeding. Naturally, the racist system that former students, the data subjects in our input data, lived in is what caused those biased forecasts, not any inherent qualities of the student.

Barocas and Selbst (2016) further detail the various ways that data, particularly "big data," can be biased. For example, because decision-making data is a compilation of past decisions, it will carry the prejudices of previous decision-makers. Furthermore, because existing societal biases influence decision-makers, the data will also reflect these biases. Occasionally, big data includes correlations that are important for decision-making but are solely based on discrimination and unfair treatment patterns.

Another issue is sample bias, which happens when a systematic inaccuracy in data collection results in a sampling of data that is not typical of the entire population, which puts the algorithm under strain. A decision-making framework that relies on this sample would inevitably exhibit bias, either towards or away from the overrepresented or underrepresented group (Barocas and Selbst, 2016; Drosou *et al.,* 2017; Chouldechova and Roth, 2018). Limitations and biases in our measurement techniques can result in biased data. The easiest scenario involves non-representative input data, which frequently results in algorithms that exhibit lower performance on under-sampled groups. As an illustration, any model that is developed (in order to recognize faces or identify features, for example) based on the automatic download of celebrity images will probably perform poorly when applied to individuals who do not belong to the categories that the celebrities represent (Buolamwini & Gebru, 2018).

The inability of several of these performance metrics to be fully satisfied simultaneously gives birth to another kind of value judgment. When discussing ethical-epistemic tradeoffs in philosophy, these kinds of value judgments are common (Gendler, 2011; Dotan, 2020). The decision is based on the users' objectives and core principles of the model, and developers are typically aware that these value judgements cannot be settled purely technically (Kearns & Roth, 2019).

Diverse discriminatory origins may arise from algorithmic bias. First, inconsistent impact can result from improperly weighted input variables into automated choices. For instance, putting too much focus on the area code in algorithms for predictive policing can result in the community of economically disadvantaged African-American neighbourhoods paired with areas of high criminality, which can then lead to the application of targeted targeting based on membership in a particular group (Christin *et al.* 2015). This is an example of indirect discrimination. Second, the choice to employ an algorithm itself may lead to discrimination. According to Diakopoulos (2015), categorisation is a type of direct discrimination in which algorithms are applied differently. Third, when specific models are applied incorrectly in various situations, algorithms may result in discrimination (Calders and Zliobaite, 2013). Fourth, biased training data can serve as both validation for the application of algorithms and proof of their efficacy in a feedback loop structure. (Calders and Zliobaite, 2013).

Besides the bias in data, biases introduced by incorrect procedures during algorithm training also contribute to unfairness. In the process of creating a (fair) algorithm, researchers must make a number of choices that could have drastically different effects on the results. These choices include choosing the dataset, choosing and encoding features from the dataset, choosing and encoding the outcome variable, being rigorous in identifying potential root cause of bias in the data, choosing and defining particular fairness standards, etc. Implicit assumptions are included in every decision that is made. This is referred to as "silent normative assumptions" by Green (2018). They remain mute, seemingly concealed by an obsession with mathematical precision and rigorous procedure during the algorithm's creation (Green, 2018b). Additionally, normative presumptions may frequently be the underlying assumptions. However, to construct a fair classifier, it is necessary to identify the characteristics and functions that characterise similarity.

Because ADS systems are created using data that consists of past human judgments, there is a problem with selective labels (Kleinberg *et al.,* 2017). Often, we only witness the outcome or designation from one perspective of the disagreement. For example, in the health field, we only track the results of patients with a particular treatment. Alternatively, when judges determine a defendant's bail, they consider only offences committed by freed offenders, not those incarcerated (Kleinberg *et al.,* 2017). It is challenging to make projections of crime rates for those incarcerated since judges may have chosen these people based on characteristics not seen in the data, leading to biased machine predictions based on observables. If researchers could see what would have happened if the people in jail had been freed, then there would be no issue. However, this is the "counterfactual" situation, which is unobservable since it never happens in the real world.

Certain types of bias are harder to classify into one of these categories since they do not always result in measurable discrimination against or unfair outcomes for protected groups. Algorithms that have an impact on our daily lives, such as text messaging autocomplete algorithms, picture search engines, and translation tools, are more likely to have this kind of prejudice. For example, Kay *et al.* (2015) show how image search results for specific terms related to vocations, such as "CEO", reflect (and even reinforce) on the prevalent stereotypes and prejudices regarding racial and gender makeup of these occupations. Most search results for "CEO" and "software developer" are often male. Google Translate provides a further example. The outcome of translating the phrases "She is

a doctor. He is a nurse." into Turkish, a language that is gender-neutral in this context, and then back to English is "He is a doctor. She is a nurse." "Representational" harms are the results of biases in algorithms that we use on a daily basis, replicating one another (Crawford, 2013). The issue with these damages is that they lead us to believe that these biased and stereotyped ideas are the standard because of how these algorithms impact the environments we encounter daily. Although the effects of this are more subtle and long-lasting, they nonetheless support the oppression and abuse of marginalised groups.

Moreover, a different kind of prejudice relates to the differentiation between algorithmic and human opinions and recommendations. People frequently think that human recommendations are superior to those made by ADS systems and value human input more highly. Furthermore, professionals that utilize ADM Systems are often subject to harsh criticism than the general public, particularly when they commit errors. Algorithmic aversion is the term for the phenomena when people refuse to use algorithmic systems even when doing so would be advantageous (Dietvorst *et al.,* 2015; 2018). The kind of decision and task at hand determine how much resistance there is to an automated system (Castelo *et al.,* 2019).

To have an in depth knowledge, we further discuss some factors that influence fairness, accountability, transparency, and explainability of algorithmic decision support.

3.2.1 *Fairness*

According to Hutchinson and Mitchell (2019), fairness is a difficult idea that different performers interpret in different ways and is dependent on the environment in which the system is used. People become more conscious of potential biases in decisions, data, and algorithm interaction when they are consistently exposed to algorithmic fairness (Brown *et al.,* 2019). They want further details on the decision-making process, that includes weighting of various parameters and whether sensitive attributes (such as gender or ethnicity) are used by the algorithm. Fairness and transparency in the development and application of algorithms, however, are becoming more and more important as decision support becomes more dependent on them. As a result, there's been a noticeable emphasis on enhancing accountability and justice in algorithmic decision support systems.

Individual fairness, collective fairness, and causality-based fairness are a few examples of typical ideas of fairness. The concept of fairness through awareness is frequently employed to achieve individual fairness, and it stipulates that similar people should be treated similarly (Dwork *et al.,* 2012). Determining the similarity function between various people is challenging, though. According to ideas of group fairness, various categories of people must be treated fairly by the algorithm. The concepts of group fairness that are most frequently applied are equal opportunity (Moritz *et al.,* 2016), calibration (Kleinberg *et al.,* 2016)., equalised odds (Hardt *et al.,* 2016), and demographic parity (Dwork *et al.,* 2012). These fairness principles are easy to understand and apply to real-world machine-learning problems. Nevertheless, their measurement features are limited to sensitive attributes and outcomes.

This means that these ideas might not be able to discern between the fair and unfair aspects of the issue. Recently, concepts of causality-based fairness have been presented to provide a more detailed

definition of fairness. Certain concepts of causality-based fairness, like Counterfactual fairness tailored to specific paths (Chiappa, 2019; Wu *et al.,* 2019; Nabi & Shpitser, 2018; Kusner *et al.,* 2017), the authors can distinguish between the sensitive attribute and the unfair causal effect when defining a causal graph among features.

A variety of technical definitions of machine learning fairness have been presented in recent years by various researchers, the majority of which codify some sort of collective justice. A commonly employed concept is the statistical parity that mandates a proportionate share of every group that must experience every potential result (Calders and Verwer 2010; Kamishima *et al.* 2011; Zemel *et al.* 2012; Feldman *et al.* 2015). Motivated by the US legal code's concept of disparate impact, recent publications have additionally investigated statistical parity approximations (Zafar *et al.,* 2015; Feldman *et al.,* 2015). Additionally, learning algorithms that penalise statistical parity violations have been established from work in these approaches (Kamishima *et al.* 2011; Calders and Verwer 2010).

A precise description of individual justice that is amenable to formalization mathematically of the Rawlsian ideal of "fair equality of opportunity" (Rawls, 1971) has recently been presented, drawing on the work of Dwork *et al.* (2012) and Joseph *et al.* (2016). According to this theory, people "should have the same perspectives of success regardless of their initial place in the social system" (e.g., ethnic background, income etc.) "who are at the same level of aptitude and have the same motivation of using it" (Rawls 1971). Therefore, this idea holds more weight than "formal equality of opportunity": Indeed, Rawls contended that a person ought to possess a practically equal chance with a different individual who has comparable natural qualities, in addition to the right to opportunity. Within their suggested methodology, Joseph *et al.* (2016) include the concept of fairness in a machine learning literature named contextual bandits sequential decision-making framework. According to their definition of fairness, the learning algorithm must never, at any stage, give preference to applicants whose qualities are inferior to those of a different candidate. Therefore, the main goal is creating an algorithm for machine learning that, while being (verifiable) fair at every stage, would (evidently) converge to an ideal conclusion. They demonstrate how machine learning algorithms can be proven to be equitable in a way that makes improving the algorithm's fairness inexpensive (in terms of the rate at which it converges to an optimal result).

A fairness metrics based on the similar concept of equality of opportunity was presented by Hardt *et al.* (2016) in an effort to accomplish two significant goals. The first step is to address the primary conceptual flaws with statistical parity as a concept of fairness. Second, in keeping with the main objective of supervised machine learning, to construct classifiers with improved accuracy. With the intention of achieving this, criterion were set for discrimination against a given sensitive attribute in supervised learning, in which the objective is to forecast a target using the features at hand.

Fairness principle(Principal fairness) is a different concept of fairness that applies to both algorithmic and human decision-making. Fairness principle integrates causality into fairness, in contrast to the current standards of statistical fairness (Hardt *et al.,* 2016; Chouldechova, 2017; Zafar *et al.,* 2017; Johndrow and Lum, 2019). Furthermore, the existing causality-based fairness requirements are not the same as principal fairness. Principal fairness, in particular, is distinct based on the counterfactual equalized odds standards in that it takes into account the decision's impact on the outcome by taking

into account joint alternative outcomes (Coston *et al.,* 2020). Furthermore, principle fairness takes into account the decision's effects rather than those of protected qualities of interest, which is different based on the counterfactual fairness standards (Kusner *et al.,* 2017; Nabi and Shpitser, 2018; Zhang and Bareinboim, 2018; Chiappa, 2019).

Fairness priciple's core tenet is that people shouldn't be treated differently from those who would be similarly impacted by a choice. Imagine a judge making the decision to hold or free a person under custody while the resolution of any felony accusations during a first appearance hearing (Imai *et al.,* 2021). Also, principal fairness is in agreement with the principle of individual fairness (Dwork *et al.,* 2012), which asserts that persons of like minds ought to be treated equally It is important to note that principle fairness evaluates similarities based on possible (factual as well as counterfactual) outcomes instead of observed variables such as resulting outcomes, covariates, or any function of them.

The three types of methods that reduce biases in the algorithms include post-processing (Hardt *et al.,* 2016), pre-processing (Feldman *et al.,* 2015; Kamiran & Calders, 2012; Wang *et al.,* 2019), in-processing (Hashimoto *et al.,* 2018; Zafar *et al.,* 2017). Zemel *et al.,* (2013) initially suggested representation learning, a popular in-processing technique. The articles make an effort to simultaneously lessen discrimination based on demographics and individual injustice. The state-of-the-art technique as of late is adversarial representation learning. This type of approach was initially put forth by Edwards and Storkey (Edwards & Storkey, 2015), who also offered a framework for reducing demographic discrimination.

3.2.2 *Accountability*

The distributed responsibility in algorithmic supply chains should be addressed by governance and accountability frameworks surrounding algorithmic systems. A given AI technology is subject to various actors' control in terms of commissioning, designing, creating, deploying, using, or monitoring. Several actors, who may not always be consistent or easy to identify, share responsibility for the operations and results of supply chains. So, generally speaking, no single actor has total control over a supply chain, even in cases when several individuals are powerful. However, the existing literature on accountability usually makes the assumption that the actors and components stay mostly constant, even while input data or models may vary. Therefore, it runs the danger of undermining the stated objectives of these mechanisms to target the wrong supply chain actors with accountability or to allocate liability to them. Accountability is a relationship in which an actor reports their actions to a forum, which then has the authority to correct the actor as necessary (Boven, 2006). Therefore, it is essential that the correct actors are paired with the right relationships for accountability systems to work.

While the concept of accountability can be conceptualised broadly, in actuality it is highly contextual. Depending on what needs to be accounted for, the ADS process will likely have multiple actors responsible for different parts, and the kinds, quantities, and arrangements of data needed for a relevant and suitable account will mostly rely on the forum for which it is accountable. (Bovens, 2006; Wieringa, 2020). Individual, hierarchical, collective, and corporate accountability are the four categories of accountability interactions that Bovens (2007) outlines according to the actor's level.

Individual responsibility entails holding each person accountable for their own actions. Stated differently, when an individual is not protected against an inquiry by their organisation or superiors. According to Boven (2007), hierarchical accountability refers to the process whereby the individuals in charge of an organisation, department, or team are held responsible for the overall outcome. The concept of collective responsibility is based on the notion that any individual inside an organisation, irrespective of their position or role, can be held responsible for the organisation as a whole (Boven, 2007). Because it is not sufficiently complex to do justice to the many variations that are significant in the assignment of guilt, shame, and blame, this type of accountability relationship is uncommon in democratic situations. When a company is held legally responsible as a non-human entity, this is referred to as corporate accountability (Bovens, 2007). For example, this is the case when discussing the terms developing firm or data controller (Vedder & Naudts, 2017; Martin, 2018).

Predictive algorithm adoption may result in algorithmic biases. Because algorithms apply values based on the optimization they were trained for, significant biases may develop if the algorithm's and the users' values differ significantly (Danks & London, 2017). When choices or policy changes are based only on predictions from an observational model, biases resulting from understanding gaps can also occur (Caruana *et al.,* 2015). Utilizing algorithmic outputs by individuals, however, will vary depending on a variety of factors, including decision focus (Green & Chen, 2019), institutional accountability frameworks (De-Arteaga *et al.,* 2020), trust (Dietvorst *et al.,* 2015), and decision context (Kleinberg *et al.,* 2018). Because of this, a biased algorithm that supports a prejudiced human being supported by an impartial algorithm, and a human who is not biased, or both of them could make unethical choices and inflict unjust harm.

In general, certain authorities have attempted to handle distributed responsibility in supply chains driven by data. Different parties may be controllers for some or all components of a chain of processing, according to rulings made frequently by the Court of Justice of the European Union (CJEU) (CJEU, 2018; 2019; Cobbe and Singh, 2021; Mahieu *et al.,* 2019). Joint controllers may be separate controllers when multiple actors' interests in the processing differ; they may be controllers when multiple players have common interests in the processing (European Union, 2016). Although acknowledging the diversity of participants in processing chains is a positive development, algorithmic supply chains may not be able to easily adapt to the more complex duty and responsibility assignments found in data protection legislation (Cobbe and Singh, 2021; Gúrses and van Hoboken, 2017; Mahieu *et al.,* 2019). According to existing understandings, suppliers of AI services may be considered data processors (the subordinate party operating merely at the direction of a controller, having restricted obligations), while clients of the service are probably data controllers (the leading party, liable for compliance and accountability). (Gúrses and van Hoboken, 2017; Mahieu *et al.,* 2019; Millard, 2021).

Roles are another way to identify actors. It might be argued that the person writing the system's specifications will be viewed differently from the system's developer or user in certain scenarios. Therefore, one could also argue that roles play a part in matching the right performer to certain parts. Three categories of actor roles can be distinguished: users, developers, and decision makers. According to Coglianese and Lehr (2017), it is crucial to consider who exactly has the authority to specify algorithms inside an organisation. Higher-level personnel are more answerable to others,

therefore they cannot be ignorant of crucial algorithmic intricacies, so there is a lot on the line when it comes to deciding who gets to make these judgements.

Due to their familiarity with design choices and special ability to imbue the algorithm with important biases and roles and obligations related to algorithmic decision-making, developers are frequently viewed as the responsible party in these situations (Martin, 2018). According to Kraemer, Van Overveld, and Peterson (2011), it is plausible to argue that software designers bear moral responsibility for the algorithms they create, as developers are unable to resist making moral decisions about what is right and wrong. As a result, developers include value judgements into the algorithmic system, either directly or implicitly. The reasoning behind this is that the user should have as much control over the selections as feasible. The creating entity bears a greater accountability burden when users are deprived of certain choices (Kraemer *et al.,* 2011; Martin, 2018).

However, this indicates that designers and/or developers should also be sufficiently sensitive to ethical issues that could result from the technology being developed (Torresen, 2018). Since decisions regarding the balancing of error rates are not typically included in specifications (Kraemer *et al.,* 2011), developers must first be able to identify and mark these ethical issues before they can consult with stakeholders as necessary and take those decisions into account.

It is imperative to pay particular attention to the users of the system and how they interact with it. Three categories of systems can generally be distinguished: human-on-the-loop, human-in-the-loop, and human-out-of-the-loop. Although this typology was initially developed for AI warfare systems, it is usefully adapted to algorithmic accountability (Citron & Pasquale, 2014; Danaherm 2016). One may say that human-in-the-loop systems enhance human practice. While these algorithms offer recommendations for potential courses of action, no action will be conducted without human approval. To describe it another way, Yeung (2017) explains these as decision-guidance processes. Human agents keep an eye on human-on-the-loop systems, but rather than having a default state of "no, unless consent is given," these systems will continue working until the human agent instructs them to stop. Lastly, some systems have human oversight removed entirely which is referred to as human-out-of-the-loop (Yeung, 2017). It may be argued that the various forms of interaction have an impact on the accounts that the user-as-actor can provide.

Therefore, it's critical to comprehend how accountability is distributed in algorithmic supply networks, including who is responsible for what for whom, what essential tasks they perform for others, who is essential to particular supply chains, and who has systemic importance.

3.2.3 *Transparency*

Transparency, which is the ability to understand a certain model is a way to promote accountability. To be more precise, transparency can be seen from the perspective of the complete model, as well as from the perspective of certain training algorithms, individual components, and parameters. Transparency is defined as the ability to consider the whole model at once in a given context. For models, low computational difficulty is, therefore, a desirable attribute. According to Lou *et al.* (2012), an alternative and less stringent definition of transparency might be that every element of

the model, such as every input, parameter, and calculation, allows for an obvious explanation. Even in the absence of the capacity to simulate a complete model or deduce the meaning of its constituent parts, a final concept of transparency may be applicable at the algorithmic level. Ananny and Crawford (2018) state that for recommendations given by algorithmic processes to be considered transparent in connection with personalized algorithms, they must be readily apparent to consumers.

There is a lot of dispute about what really qualifies as a transparent explanation, how transparent something should be, who should know, and why (Sloan & Warner, 2017). In regard to this, Shin and Park (2019) describe algorithmic transparency as the need for consumers to be able to comprehend the process by which an artificial intelligence system arrives at a conclusion or forecast. According to Diakopoulos and Koliska (2016), Transparency demands that both the inputs to the algorithm and the algorithm itself be accessible and understandable. Algorithmic transparency is associated with concepts like interpretability, explainability, and algorithmic visibility. This makes the choices made in algorithm outputs interpretable and allows computational procedures and intentions to be fully taken into account (Meijer, 2014). For instance, transparency serves as a means of assessing the validity and justification of the algorithm's purpose, the use of authentic data, and the implementation of an algorithm that is both statistically and mathematically acceptable for the task at hand (Courtois & Timmermans, 2018). People tends to use content appropriately and to trust algorithm and the outputs of the generated material when they understand how algorithms operate and how machine learning operations are carried out (Shin *et al.,* 2020).

It has been determined that a variety of strategies lead to increased openness in algorithmic decision making. Repeated interactions with a system can make users aware of an algorithm (Rader and Gray, 2015). Users may encounter unexpected or puzzling information that contradicts their expectations (Rader, 2017; De Vito *et al.,* 2017), prompting concerns about potential algorithmic bias (Eslami *et al.,* 2017). Alternatively, users might be motivated to gain a deeper understanding of computational outputs to devise strategies to mitigate unfavorable outcomes. (Lee *et al.,* 2015). But this "organic" awareness is neither uniformly distributed across users nor methodical.

Algorithmic audits represent another form of transparency method aimed at examining the impacts and functioning of algorithmic decision-making systems (Mittelstadt, 2016). As noted by Sandvig et al. (2014), algorithm audits can operate at various levels, each offering different degrees of accountability and visibility. However, given the restrictions often imposed by system providers in their terms of service, audits typically require independent conduction. Some have suggested that platforms intentionally withhold operational information to safeguard themselves against competitors or individuals attempting to manipulate the system (Burrell, 2016).

According to Sandvig *et al.* (2014), an auditing technique views the decision-making process as a "black box," with visible inputs and outputs but hidden internal workings. But as numerous studies have demonstrated, analyzing decision processes and systems in a black-box is the least effective way to comprehend how they behave (Datta *et al.* 2015). Using their AdFisher technology, Datta *et al.* (2015) explore the opacity, or lack of transparency, in web-based advertisements. To delve deeper into the transparency provided by Google's Ad Settings, they conducted a number of experiments. Specifically, they examine whether accessing websites associated with a certain

interest might result in a modification of the shown ads that is not recorded in the settings. Their technique failed to display any profiling, but they did find instances of opacity in cases when there were notable changes in the ads displayed to different profiles.

Giving explanations is a popular strategy in recommender systems (Tintarev and Masthoff, 2011), which is a third kind of method to increase openness and potentially address issues brought on by opaque algorithmic decision-making systems (Lee *et al.,* 2015).

3.2.4 *Explainability*

The degree to which an instance's feature values are connected to its model prediction in a way that makes sense to people is known as explainability (Rai, 2020). Transparency makes AI easy to understand and allows algorithms to be discussed in terms of how specific outcomes are achieved. Algorithmic AI systems propose or suggest actions based on opaque processes that are incomprehensible to the general public (Renijith *et al.,* 2020).

Although explainability seems like a good idea, it might be challenging to implement. According to Belle and Papantonis (2021), there are four ways to increase explainability: employing graphical visualization techniques, explaining an instance rather than a generalization, explaining an explanation by simplifying the importance of each aspect to the decisions, and providing explanations by simplicity. They also talk about how difficult it would be to implement such recommendations at the same time. Simple explanations may not be accurate, features may be connected, local explanations may fall short of giving the whole picture, and graphical representations may rely on assumptions about data that are not always true. It is believed that explainability will increase openness and confidence in AI. Situational circumstances also impact trust, however they may do so in unexpected ways (Bannister & Connolly, 2011).

According to Simkute *et al.* (2021), comprehending how people engage with algorithms and what data they require to support their decision-making processes is crucial for explainability to be effective. Explanations should be designed to accommodate the distinct decision-making and sense-making techniques of both industry experts and beginners in order to guarantee that they can assist decision-makers in retaining meaningful agency. To date, only a few studies (De-Arteaga, Fogliato Chouldechova, & 2020; Green Chen, 2019) have attempted to investigate the interactions between humans and algorithms in a decision-making context, and even fewer have looked at the factors influencing human decision-making and sensemaking strategies (Simkute, Luger, Evans Jones, 2020). In addition, there aren't many design guidelines that indicate which explainability technique would be best in a given scenario, taking into account contextual circumstances, the decision maker's needs, and variations in human reasoning. This is true despite the abundance of explainability techniques accessible. Guidelines outlining how explainability could be incorporated into currently used applications that are utilized in real-world circumstances are also lacking (Eiband *et al.,* 2018). These guidelines should include information on what should be explained, how to present explanations in the system, and how to take real-world constraints into account.

Another method for promoting transparency in algorithmic decision-making is presenting human-interpretable explanations of the decision-making processes. In a recent study, Ribeiro *et al.* (2016)

suggested that in order to address the so-called "trusting a prediction" problem, one should (i) explain each prediction individually and (ii) choose a number of these explanations and predictions to address the so-called "trusting the model" problem. There are several technical methods that can enhance an algorithm's explainability and interpretability such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations).SHAP values provide a unified measure of feature importance, while LIME constructs a locally interpretable model around a prediction, enabling explanations for classifier's predictions (Salih *et al.,* 2023). Additionally, they suggested a non-redundant manner of explaining models by displaying exemplary individual forecasts together with their justifications. The value of these explanations was demonstrated in the paper through simulated and human subject experiments on a variety of scenarios, including (i) determining whether to trust a prediction, (ii) identifying a classifier that lacks reliability and should not be trusted, and (iii) selecting between various classification models, among other tasks.

3.5 Guidelines for Ethical Design and Implementation.

The standards expected of those who use algorithms are still largely undefined and unevaluated, despite the fact that integrating ethical values (most notably, justice, accountability, and transparency) into algorithms has received much attention. Since precision is not the only quality of morally and responsibly made decisions, appeals to embrace machine learning models often frequently center on this aspect (Corbett-Davies *et al.,* 2017; Kleinberg *et al.,* 2017). For example, among other requirements, decisions must be ethical, precise, equitable, consistent, and able to be corrected in accordance with the procedural justice principle (Levanthal *et al.,* 1980). Although algorithms have the potential to increase prediction accuracy, many of these principles are not well served by them because of their limited capacity for reflexive thought and adaptation to new or unusual situations (Alkhatib & Bernstein, 2019). Because of this, organizations that use algorithmic counsel may praise the algorithm for its ability to yield insightful data while also warning that the program shouldn't be used to make decisions (Wisconsin Supreme Court, 2016).

In the context of using machine learning models to help make predictions (or decisions based on predictions), Green & Chen (2019) propose three desirable behavioural principles. The following are their three guiding principles:

Accuracy: Predictions made by those utilizing the algorithm ought to be more accurate than they would have been otherwise.

Reliability: Users ought to adjust; they should appropriately evaluate both their own and the algorithm's performance, taking into account the algorithm's errors and accuracy in the process.

Fairness: When interacting with the system, users should act impartially toward sensitive characteristics like gender and ethnicity.

The amount of frameworks, moral standards, and design and development principles for algorithms has increased dramatically in recent years. Google and Microsoft are two examples of companies that have released standards for the creation of morally sound artificial intelligence and algorithms. The European Commission (2020) suggested a rule in the European Union to incorporate the

principles of safety, privacy, responsibility, human oversight, and non-discrimination into algorithmic systems.

In order to guarantee that AI systems be created, implemented, and utilized in a reliable manner, we outlines four ethical guidelines that are based on fundamental rights.

The principle of respect for human autonomy

The goal of the fundamental rights on which the EU is based is to guarantee that people's freedom and autonomy are respected. When engaging with AI systems, humans must be able to maintain complete and effective control over their own lives as well as participate in democratic processes. Artificial intelligence (AI) systems should not force, control, subjugate, trick, or herd people in an unwarranted manner. Rather, their design ought to enhance, complement, and bolster human cognitive, social, and cultural capacities. Adhering to human-centric design principles is imperative when delegating tasks to AI systems, while also allowing ample space for human decision-making.. This entails providing human supervision over AI systems' work operations.

The principle of prevention of harm

Artificial intelligence systems ought not to injure, worsen, or have any other negative effects on people. This includes safeguarding people's mental and physical integrity in addition to their dignity. Both AI systems and the settings in which they function need to be safe and secure. They have to be strong technically and made sure they can't be used maliciously. People who are vulnerable ought to be given more consideration and involved in the creation, application, and usage of AI systems. Instances characterized by power or knowledge imbalances, such as those observed between employers and employees, businesses and consumers, or governments and citizens, might result in or worsen negative outcomes owing to AI systems, also require special attention. Taking into account the natural surroundings and all living things is another aspect of preventing injury.

The principle of fairness

AI systems must be developed, implemented, and used fairly. Fairness could have both a substantive and a procedural dimension, even if it was recognize that there are many diverse interpretations of what constitutes fairness. The substantive dimension suggests a commitment to: making sure that expenses and benefits are distributed fairly; and making sure that people and groups are not subjected to unjust prejudice, discrimination, or stigmatization. Artificial intelligence systems have the potential to improve society justice by preventing unjust prejudices. Equality should be promoted with regard to access to technology, goods, services, and education. Furthermore, people's freedom of choice should never be unjustly restricted or deceived as a result of using AI systems. Furthermore, fairness requires artificial intelligence practitioners to carefully weigh competing interests and goals and adhere to the proportionality concept between means and ends. The ability to challenge and seek meaningful redress against judgments made by AI systems and the humans who operate them constitutes the procedural dimension of fairness. To achieve this, the decision-making procedures must be understandable and the entity responsible for the choice must be recognizable.

The principle of explicability

Users' trust in AI systems must be established and maintained through explainability. This means that procedures must be clear, AI systems' purposes and capabilities must be freely disclosed, and judgments must, to the greatest extent feasible, be explicable to all parties involved, both directly and indirectly. One cannot properly contest a decision in the absence of such facts. It is not always feasible to explain why a model produced a specific output or result, or what combination of input factors led to that. These situations are known as "black box" algorithms, and they call for extra care. Under these circumstances, when the system as a whole respects basic rights, additional measures to enhance explainability (such as traceability, auditability, and transparent communication regarding system capabilities) may become essential. The setting and the seriousness of the repercussions in the event that the output is incorrect or incomplete determine how much explicability is required.

Hence, AI developers are urged to place greater emphasis on the social and ethical dimensions of constructing equitable AI, actively striving to eradicate bias from the AI models they create (Sullivan & Fosso Wamba, 2022). However, the involvement of AI developers, their affiliated organizations, and policymakers in mitigating AI bias through equitable processes remains largely unexplored. Despite extensive research on the significance of explainable AI, which focuses on developing AI that is transparent and comprehensible to society, (Samek *et al.,* 2019; Colander, 2022; Hunkenschroer & Luetge, 2022). There are also considerable amount of research as regards Accountability. To ensure better processing and effective strategies for developing fair algorithmic decision support, understanding the role the organisation plays is imperative.

4.0 RESULT AND DISCUSSION

This chapter reviews, synthesizes, and compiles the various literature selected for the study. The findings are classified into themes to give a clear illustration of the findings on organizational factors that influence fairness in Algorithmic decision support. The organisations are viewed from two angles: from the angle of organisations designing the algorithm and from the angle of the clients organisations (Stakeholders) adopting/making use of algorithmic decision support in their institutional process and procedures.

This systematic literature review has the objective to identify factors that influence fairness in algorithmic decision-support, and to identify recommendations and suggestions for future research in those articles. Both relevant articles on algorithmic decision-support and Machine Learning decision-support were reviewed. The finding show several organizational factors that influence fairness from the employee level (Practitioners) and policy makers.

To address the research question – "What are the organizational factors that influence fairness in algorithmic decision support?" – the search was conducted in three main databases, as well as a grey search using pearl sampling/Snowballing for additional growing through citation, after the application of the inclusion and exclusion criteria, twenty-three (24) articles were reviewed Ten (10) empirical articles, Eleven (11) literature Reviews and three (3) Systematic Literature Review. Figure 2 shows the documentation process that resulted in the selected articles. Figure 3 shows the publication year of the articles reviewed.

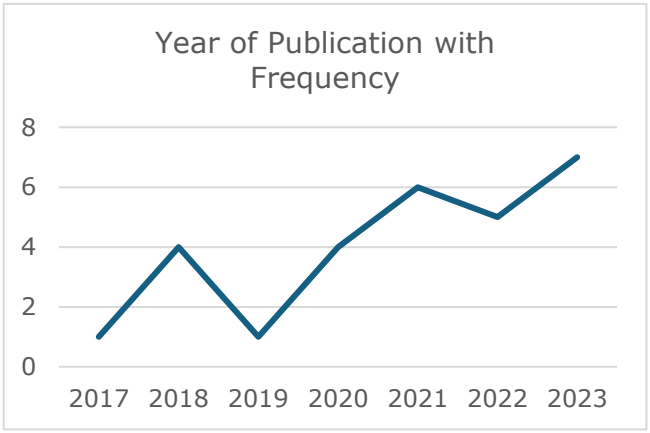

Figure 3: The distributions of the articles in years

The organizational factors identified cut across different phases of the algorithmic model processing which can be classified into three groups: Preprocessing, In-processing, and post-processing. Various factors were discovered from the different articles researched. Figure 4. shows the different factors identified and how often they occur in the articles.

Figure 4: Graph showing different factors identified from the selected articles

Based on the literature reviewed, a governance structure (Implementation based) was developed to delineate the interrelations of responsibilities within organizations. This structure is categorized into three main areas, representing the implementation of key factors within the organization. These categories provide a comprehensive framework to ensure effective governance and accountability in ensuring fairness algorithmic decision-support system design.

1. **Leadership-related:** This category encompasses factors related to leadership style, practices, or characteristics. Individuals within this category may include Product Managers, Project Managers, or other leaders responsible for guiding and directing teams.
2. **Task-related**: This category includes factors related to specific duties and tasks within the organization. Individuals in this category, such as practitioners or team members, are responsible for carrying out these tasks and responsibilities.
3. **Organizational-related:** This category encompasses factors related to organizational culture, policies, procedures, and resource allocation. Individuals in this category may include higher-level executives, HR personnel, or those responsible for shaping and implementing organizational practices and structures.

Figure 5. Shows the pictorial illustration of the factors base on implementation strategy.

Figure 5: Governance Structure reflecting interrelationship in Organisation Responsibility for fairness in ADS. (Implementation of Factors).

The role and responsibility cut across different units as it is everybody's duty to work together to achieve fairness. Table 4. Show the distribution of the factors across different units that could be responsible for implementing them. It is essential to identify the necessary stakeholders. The list of stakeholders, adapted from Adensamer *et al.,* (2021), includes: Organizational Management Board, Governmental agencies making decisions on the introduction of ADS, Individuals responsible for the integrating ADS systems into organizational workflows, Developers creating or modifying systems, Quality assurance personnel and Auditors. These stakeholders play crucial roles in the implementation, evaluation, and oversight of Algorithmic Decision Systems (ADS). Identifying the stakeholders provides insight into who is responsible for driving fairness at each stage of the process (Curto *et al.,* 2023; Lepri *et al.,* 2018; Ueda *et al.,* 2023). This knowledge helps in understanding the roles and responsibilities crucial for ensuring fairness in the design, implementation and operation of Algorithmic Decision Support systems.

The Summary of the implementation and correlation to the theme is shown in Table 3.

Table 3: Bridging the classifications of the theme in relation to the implementation.

	Organization	Leadership	Task
Governance	✓	✓	
Social Responsibility	✓	✓	✓
Technical			✓
Training & Development			

Table 4: Factors Across Different Units Responsible for Implementation

	Leadership related	Task related	Organisation related
Ethics policy	✓	✓	✓
Auditing tools		✓	✓
Resource allocation	✓		✓
Leadership practices and commitment	✓		✓
Collaboration		✓	✓
Interdisciplinary team		✓	✓
Data Management		✓	
Knowledge Management		✓	✓
Communication	✓	✓	✓
Explainability and Transparency	✓	✓	✓
Problem identification and Problem solving skills		✓	✓
Accountability	✓	✓	✓

The identified factors were categorized into four themes: Governance, Social Responsibility, Technical, and Training & Development. These themes, illustrated in Figure 4, provide an outline of the factors influencing fairness in algorithmic decision-making. Various terminologies from different articles were harmonized, merging similar factors under unified names and categories for clarity and consistency. These consolidated factors are organized within relevant themes, as illustrated in the accompanying Table 3. Identified factors such as Quality Data, Representative Data, Data Review, Context Specific, Data Accuracy, Socio-technical Perspective were all merged together as Data Management. AI ethics training, Diverse Data Training, Emotional Intelligence Training, Education were all grouped together as **Knowledge Management**. Collaboration (Diverse Team/Interdisciplinary Team), Data Management (Data Review, Checklist for data collection, context specific data, data accuracy, data quality, fair data representation, increase data collection, representative data, quality data set), Ethical policy (Policies and Governance), Fairness Metrics (Evaluation metrics, domain specific metrics, checklist), Human judgement (Human-in-the-loop/Human agency/Human oversight/integration of Human Factor), Organizational framework (Organization structure/Culture/processes/system), Performance metrics (Impact assessment),Problem solving skill (Addressing and detecting error, domain specific problems, manage fairness in each phase) Procedural Justice/fairness, Resource allocation/Responsible organization.

Table 5 : Themes - Organizational Factors Influencing Fairness in Algorithmic Decision Support

THEMES	FACTORS
Governance	Ethics Policies Audit Resource Allocation (Responsible distribution) Leadership and commitment Fairness Metrics Organisational Framework (Responsible Organization) Managing Stakeholders Expectation Performance metrics
Social Responsibility	Collaboration Human Judgement Sociotechnical approach
Technical (Model Building)	Data Management Problem solving skills Procedural fairness Implementation of Procedure Feature Selection Model Testing Model Review Monitoring and assessment Trade-off New methodology approach (Substantive)
Training and development	Knowledge Management Communication Explainability Transparency Accountability

4.1 Organizational Factors influencing Fairness base on Theme

4.1. Governance

A comprehensive approach to governance encompasses proactive measures to uphold ethical standards, promote transparency, and address disparities in resource allocation. Enhancing governance in organizations involves a multifaceted approach that integrates various factors to ensure fairness, accountability, and ethical practices throughout all levels of operation. Governance encompasses a broad spectrum of elements vital for fostering ethical conduct, transparency, and equitable resource distribution within organizations. Governance can be further grouped in to two sessions metrics and fairness, and secondly management practices. The grouping is illustrated in Figure 6.

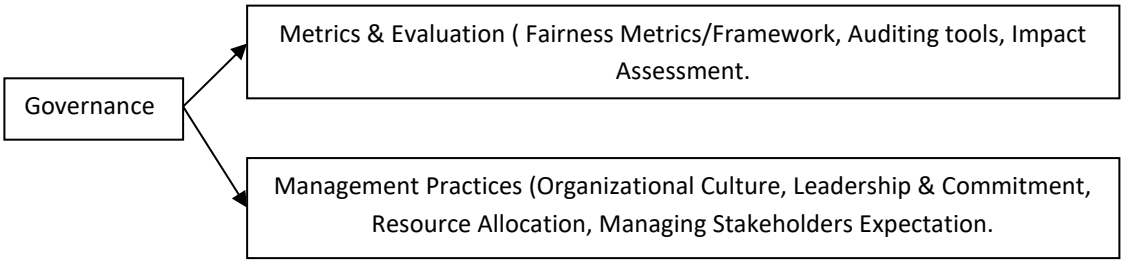

Figure 6 : Theme – Division of Governance Theme

4.11. Metrics and Evaluation: This has to do with the assessment tools to improve fairness in algorithmic decision support; Ethics policy, Fairness metrics, Auditing and Impact assessment. Ethics Policy serves as the cornerstone of organizational conduct, outlining principles and guidelines for ethical decision-making and behaviour. Ethical principles should guide the development and deployment of algorithmic systems to ensure that they do not perpetuate or exacerbate unfairness or discrimination. Provision of Ethical ranking from ethics committee (Raji *et al.,* 2020; Shneiderman 2021; Ueda *et al.,* 2023; Veale *et al.,* 2018; Xivuri *et al.,* 2023).

Fairness Metrics are metrics such as checklists and frameworks that help organizations assess and measure fairness in their policies, practices, and decision-making processes. These metrics provide valuable insights for addressing potential biases and disparities, Developer highlight integrating fairness checklists into organisational goals to ensure adaptation right from the start (Ferrara *et al.* 2023; Green 2022; Lepri *et al.* 2017; Madaio *et al.,* 2020). Researchers like Ferrara et al. 2023; Lee and Singh 2021; Rana *et al.,* 2023; Srinivasan and Chander 2021, make mention of some fairness metrics in their research that could help in achieving fairness. However, they need to be used within context for effective result.

Table 6: Fairness tools identified from the selected literatures

S/N	Fairness tools	References
1.	IBM"s AI Fairness 360	Ferrara et al. 2023; Lee and Singh 2021; Rana *et al.,* 2023; Srinivasan and Chander 2021)
2.	Microsoft Fairlearn	Ferrara et al. 2023; Lee & Singh 2021; Rana *et al.,* 2023
3.	The Aequitas tool	Ferrara et al. 2023; Lee & Singh 2021; Rana *et al.,* 2023
4.	Themis-ML	Ferrara et al. 2023.
5.	Scikit-fairness/Scikit-lego	Lee & Singh 2021; Rana *et al.,* 2023
6.	PyMetrics Audit AI	Lee & Singh 2021; Rana *et al.,* 2023
7.	Google What-if-tool	Lee & Singh 2021; Rana *et al.,* 2023
8.	ML Fairness Gym	Rana *et al.,* 2023

Auditing Tools are Implementing tools that enable organizations to effectively monitor compliance with ethical standards and identify areas for improvement or corrective action. Auditing refers to the process of systematically evaluating and validating the integrity, fairness, and adherence to established standards within an organization's processes or systems. Auditing in the context of Algorithmic Decision Support (ADS) systems refers to the processes and practices developed to ensure the transparency, fairness, and accountability of algorithms. (Schneider 2021; Xivuri *et al.,* 2023; Green 2022). These audits are essential for detecting biases, errors, and potential harms, ensuring that the algorithm's design aligns with ethical standards. Auditing Algorithmic Decision Support (ADS) systems involves a comprehensive approach that addresses not only technical accuracy but also ethical considerations, privacy concerns, and the broader social impact of these systems. Auditing encompasses both internal and external evaluations, aiming to identify and address potential risks, gaps, or non-compliance with regulations or ethical norms. It provides assurance to stakeholders, promotes procedural justice, and drives organizational improvements toward alignment with ethical standards and regulatory requirements (Raji *et al* 2020).

Performance metrics (Impact assessment), a crucial component of auditing, involves evaluating the effects and consequences of algorithmic systems on various stakeholders and societal contexts. It entails analyzing the impacts of algorithmic decision-making processes, such as those used in ADS, on individuals, communities, and broader social structures. Factors considered in impact assessments include fairness, transparency, accountability, and potential harms resulting from the deployment of algorithmic systems. Impact assessment aims to ensure that algorithmic systems are deployed with due consideration for their potential effects on users and society at large (Corbett-Davies *et al.* 2017; Metcalf *et al.,* 2021; Veale *et al.,* 2018.). By addressing these factors, organizations can ensure that their ADS systems are fair and responsible.

Developing tools that provide real-time feedback and suggestions, as well as asking the questions as highlighted by Koefer *et al.* (2023), ensures that fairness considerations are incorporated throughout the data pipeline. Simulation tools can prototype and simulate user-system interactions, allowing developers to anticipate sensitive contexts and potential fairness issues. These simulations can help identify risky conversation patterns or harmful forms of personalization. Performance metrics for algorithmic models can be complex and value-laden, making it challenging to communicate effectively with stakeholders.

4.12. Management Practices

This includes the Organisational culture, Leadership and Commitment and resource allocation and Managing stakeholders expectation. Organizational Culture has to do with cultivating a culture of responsibility, inclusivity, and stakeholder engagement. This is paramount for upholding ethical standards and mitigating adverse impacts.

Resource Allocation (Responsible Distribution) ensures fair and responsible allocation of resources which is essential for ensuring equitable access to opportunities and capabilities within the organization. Kochling *et al.* (2020); Ueda *et al.,* (2023); Veale *et al.,* (2018) Resource allocation entails considering factors such as need, merit, and impact when distributing resources. Economic disparities among organizations can significantly influence their capacity to invest in novel

technologies and practices. Larger, more affluent organizations often possess greater resources to explore innovative ideas and technologies, while smaller or financially constrained organizations may resort to adopting and adapting practices from others.

Ensuring effective resource allocation is paramount, as organizations must address resource disparities to prevent the perpetuation of systemic inequities. Veale *et al.,* (2018) observed that scaling up social practices associated with algorithmic decision-support systems can present challenges, as they may not be as readily transferable as the software itself. Introducing new models or practices may necessitate substantial investments in training and process transformation, posing financial hurdles for smaller organizations. Larger vendors may offer pre-trained models to less-resourced organizations, yet this can potentially result in issues regarding model transferability and effectiveness. Organizations procuring these models may lack the in-house expertise required to comprehend, customize, or enhance them to suit their specific requirements. Adensamer *et al,.* (2021) ADS often operate under context-specific assumptions, pertaining to both the problems they aim to address and the expertise necessary for their effective utilization. This contextuality can pose challenges when attempting to scale up practices across diverse organizations or regions, necessitating careful consideration of local contexts. To ensure a fair and efficacious implementation, it is imperative to consider the broader socio-economic context and the distinct needs of organizations when sharing system models, and to ensure adequate resources are allocated for equitable implementation.

Leadership and Commitment are pivotal in championing ethical practices and fostering fairness and accountability throughout the organisation. Their commitment to promoting fairness and transparency sets the tone for the entire organization (Shneiderman 2021). Organizational leaders bear the responsibility of providing developers with the necessary tools and resources to design and deploy systems that prioritize fairness and accountability. Organisations need to put structure in place that help practitioners in negotiating with other teams within the organisation (Cramer *et al* 2018). Integrating fairness metrics into organizational goals ensures that ethical considerations are embedded into the fabric of decision-making processes. It is important that organizations resist the temptation to prioritize speed over fairness in the development of algorithmic Decision-support (ADS). As highlighted by Madio (2020), the pursuit of expediency should not come at the expense of ethical integrity.

Adensamer et al. (2021) noted that sharing responsibility can help reduce bias in organizations. To facilitate this, they developed a responsibility distribution tool called VERA, which aids in effectively allocating responsibilities and thus mitigating bias. Similarly, Curto et al. (2023) emphasized the importance of identifying stakeholders and ensuring their agreement on fairness objectives, aligning with (Adensamer et al. 2021)'s findings. Fairness initiatives may face resistance or lack of support from team or company leadership. Overcoming organizational barriers is essential for successfully improving fairness in Algorithmic Decision-Support Systems.

4.2. Social Responsibility

Social Responsibility involves the ethical obligations to act in a manner that benefits the society as a whole. It involves considering the impact of one’s action on various stakeholders. Corbett-Davies

et al., (2017), observed that the immediate utility of a decision rule may not accurately reflect its long-term costs and benefits, underscoring the significance of considering broader societal impacts and historical inequalities in evaluating fairness (Aysolmaz *et al* 2023). Social Responsibility involves Collaboration, Interdisciplinary Teams, and the incorporation of Human Judgment. Collaboration facilitates coordinated efforts among stakeholders to address fairness considerations effectively throughout the development and deployment of ADS systems. (Ferrara *et al.,* 2023; Holstein *et al.* 2019; Lepri *et al.* 2017) Engaging experts from diverse domains ensures that ethical standards and fairness goals are integrated into the ADS pipeline. By harnessing collective expertise, collaboration promotes fairness in Algorithmic Decision-support systems and addresses ethical challenges comprehensively (Cramer *et al.,* 2018). Action research, characterized by collaborative endeavours between researchers and practitioners to tackle real-world issues, presents a valuable methodology for promoting fairness in algorithmic decision-making.

Promoting cross-team knowledge sharing to identify blind spots can deepen the understanding of fairness issues (Holstein *et al,* 2019). Interdisciplinary teams further enhance fairness and reduce bias in AI systems. By bringing together individuals with different perspectives and backgrounds, these teams ensure that a wide range of viewpoints is considered during development. The diversity and representation of decision-makers and stakeholders involved in the development and deployment of algorithmic systems can influence fairness outcomes. Ueda *et al.,* (2023); Xivuri *et al.,* (2023) Organizations that include diverse perspectives and experiences are better equipped to identify and address potential biases and inequalities. This diversity fosters inclusive practices and helps create fair ADS.

Veale *et al.* (2018) identified over-reliance or under-reliance on decision support as a cause for unfairness in AI systems. To address this issue, they proposed introducing human judgment to exercise discretion in decision-making. This notion was echoed by other researchers like Corbett-Davies *et al.* (2017), who emphasized the need for consistent application of discretion without introducing bias. While algorithms and decision rules provide a structured approach to decision-making, acknowledging the significance of discretionary assessment for individual cases is crucial. However, ensuring consistency in the application of discretion is vital to prevent bias. Thus, achieving fairness necessitates finding a delicate equilibrium between automated decision-making and human judgment. This underscores the importance of integrating human oversight into decision-support implementation and interpretation processes. Expanding on this concept, Starke *et al.* (2022) defined the "human-in-the-loop" approach, wherein humans are directly involved in the decision-making process of automated systems.

Introducing Human Judgment, also known as the "human-in-the-loop" approach, is essential for ensuring fairness and accountability in decision-making processes (Lepri *et al.,* 2017; Kochling *et al.,* 2020; Koefer *et al.,* 2023; Rana *et al.,* 2023; Holstein *et al.,* 2019). This approach allows humans to intervene and make critical judgments at any stage of an automated decision-making system's operation (Ueda *et al.,* 2022; Pessach 2022). It ensures that decisions consider complex factors not captured by the automated system alone, maintaining human oversight and aligning with broader societal values. Aysolmaz *et al.* (2023) also noted that incorporating a Human-in-the-Loop (HITL) approach can help address transparency concerns. However, Hemann (2022) argues that human

judgment can introduce the biases we aim to avoid. While human judgment can be beneficial, it must be applied with caution, discretion, and proper guidance to prevent creating more challenges.

To clarify further:

- Human-in-the-loop approach: Humans are actively involved at every stage of decision-making in automated systems, ensuring decisions consider complex factors. While the system provides guidance, no action is taken without human consent.
- Human-on-the-loop approach: Humans intervene during system design and monitor its operation, overseeing the system and intervening as necessary. Tasks proceed unless halted by humans, allowing for ongoing monitoring and adjustment.
- Human-in-command approach: Humans have authority over the system's overall impact, overseeing its economic, societal, legal, and ethical implications. This grants humans control over the system's deployment and alignment with broader values.
- Human-out-of-the-loop systems lack human oversight entirely, posing potential risks and limitations.

The studies reviewed have highlighted the intricate nature of fairness in algorithmic decision-making, emphasizing the need to consider socio-contextual and socio-technical factors to address its limitations effectively. There is also the need to consider groups and subgroups, as it is impossible to simultaneously satisfy the need for fairness for all groups. Putting things into perspective within the right context will help in having a fair system this aligns with (Mehrabi *et al.,* 2021) study. This recognition highlights the importance of integrating a broader understanding of societal dynamics and technological constraints into the design process. Incorporating socio-contextual factors allows for the inclusion of human judgment (human-in-the-loop), which is essential for overseeing algorithmic decision-support systems (ADS) to mitigate errors. Empowering users to exercise discretion can significantly enhance the efficiency of ADS.

The sociotechnical perspective emphasizes incorporating social context into ADS design, going beyond technical factors like transparency and accuracy to achieve fairness. It suggests that fairness in decision support systems requires acknowledging tensions, facilitating open discussions, and implementing safety measures.

4.3. Technical (Model Building)

The impact of algorithms on fairness hinges greatly on their design and application (Curto *et al.,* 2023). While algorithms hold promise for enhancing efficiency and equity, their design and deployment presents intricate challenges that necessitate thorough examination by researchers and policymakers. Key technical factors influencing fairness include Data Management, error detection and mitigation, implementation procedures, feature selection, model testing, and review. Data Management stands as a cornerstone in ensuring fairness. Evolving data collection practices can alter data distributions, thereby impacting algorithmic models and their performance. Understanding and managing these shifts are essential for upholding fairness and accountability. As

noted by Holstein *et al.* (2019), organizations must support practitioners in conducting fairness-aware data collection and curation.

In the context of the justice system, fairness challenges are prevalent. Data limitations hinder the accuracy of risk assessment models like COMPAS, resulting in disparities in decision-making outcomes. Furthermore, algorithmic decision-making raises ethical concerns, as biases within the data or algorithms can perpetuate systemic inequalities. Detecting discrimination within algorithms, such as COMPAS, is complex and requires careful scrutiny. Concerns arise regarding the potentially discriminatory nature of risk scores, whether by design or oversight. Evaluating score calibration and demographic disparities is crucial for ensuring fairness and equity in algorithmic decision-making.

Ueda *et al.,* (2023) research on artificial intelligence in healthcare observed that ensuring that the data used for ADS development and training are diverse and representative of the target population. This involves collecting data from a wide range of sources to accurately reflect demographics, characteristics, and potential disparities. Incorporating data from various populations, age groups, cultural backgrounds, and settings helps prevent biases from occurring in ADS systems. The quality and representativeness of data used for training significantly influence the development of fair models (Corbett-Davies *et al.,* 2017, Curto *et al.,* 2023). Therefore, teams require assistance in collecting and curating data to ensure that fairness considerations are adequately addressed, ultimately improving the fairness of datasets. Practitioners acknowledge the significance of integrating fairness, particularly during the Data Analysis and Dataset Experimentation phase of the Algorithmic Decision-support lifecycle. This phase is considered pivotal for addressing fairness concerns, as it encompasses key steps such as data collection, preprocessing, and exploration, where biases may arise or be magnified.

Trade-off: The finding shows that One of the central dilemmas faced by developers is balancing fairness with accuracy. While accuracy is crucial for effective decision-making, prioritizing fairness is equally essential to mitigate biases and promote equity, making it challenging to address both simultaneously (Ferrara *et al.,* 2023; Martins, 2021; Green, 2022; Pessach, 2022). Striking the right balance between these competing objectives often requires making trade-offs that optimize both fairness and accuracy. To achieve algorithmic fairness, Green (2022) suggests prioritizing an increase in prediction accuracy to ensure decisions are based on accurate judgments about individuals. However, (Ferrara *et al.* 2023; Martins 2021), highlighted the challenge that it's often difficult to achieve both accuracy and fairness simultaneously. Addressing this concern, Green (2022) proposed a formalism response, which involves balancing the trade-offs between competing metrics. Considering the complexities of balancing accuracy and fairness, efforts to enhance prediction accuracy in algorithmic decision support can also promote fairness by carefully integrating these two goals. Achieving this balance requires a thorough understanding of the model being developed and careful consideration of stakeholders' expectations. Balancing these trade-offs is essential for developing a fair and effective AI-driven decision-making system.

Feature selection: Corbett-Davies *et al.,* 2017; Ferrara *et al.,* 2023; Lepri *et al.,* 2018; Park *et al.,* 2022; Rana *et al.,* 2023; Srinivasan and Chander 2021; Veale *et al.,* 2018; all weigh in on

feature selection, its challenge and how careful selection of features could help influence fairness. While including certain features could contribute to a fairer system, some organizations choose to exclude these features to avoid legal issues. Sensitive characteristics are often excluded for this reason. Ferrara et al. (2023) highlighted that sharing sensitive medical data could improve fairness, but this raises privacy concerns. Practitioners are thus faced with the trade-off between fairness and accuracy. Organizations need to strike a balance, incorporating useful data while navigating potential legal pitfalls.

Problem-Solving Skills (Detection and Addressing Errors): Algorithmic decision support often encounters obstacles that impede its ability to achieve the desired level of fairness. While research endeavours have focused on mitigating these challenges and improving system performance, it is crucial to consider the perspectives of both the organizations deploying the system and those utilizing it for decision support and this was supported by (Cramer *et al.,* 2018). Identifying the root causes of unfairness is paramount in fostering the development of a fair and just algorithmic decision-support system.

Various factors contribute to bias and unfairness, such as complacency, automation bias, and prioritizing non-functional aspects like accuracy and security over fairness as seen from the articles reviewed. Additionally, blind spots among developers and delayed error detection exacerbate these issues, (Koefer *et al.,* 2023) suggests that planning, building, deploying, and monitoring are crucial steps that can help developers avoid blind spots. Complacency among users, particularly within organizations utilizing decision support systems, can lead to detrimental outcomes. Users may become overly reliant on Algorithmic Decision Support (ADS), resulting in errors when interventions are delayed. Automation bias poses another significant challenge, wherein operators accept automated support without critical assessment. Overreliance on ADS systems can amplify this bias, compromising decision-making processes. Error detection and resolution are critical but often neglected aspects of fairness. Developers may remain unaware of issues until deployment, potentially perpetuating biases inadvertently through their resolution methods. To address these challenges, a deeper understanding of discretion and the enhancement of model outputs is essential.

Curto et al. (2023) suggest training developers to create models capable of identifying errors. Early detection of errors before deployment is crucial as it helps address issues before they escalate, significantly improving overall system reliability and fairness. Additionally, implementing mechanisms for detecting biases and unfairness, such as tools for proactive monitoring and feedback collection helps organization in achieving a fair ADS system.

Monitoring and Assessment: Regular monitoring and assessment of the system are essential for making necessary adjustments to ensure optimal performance and fairness using necessary tools (Curto and Comim 2023).

New Methodology (Substantive): Green (2022) suggests moving beyond formal fairness systems that rely solely on mathematical formulas, advocating for a new methodology that incorporates contextual factors, particularly in the justice system. Substantive algorithmic fairness aims to balance fairness metrics within risk assessment tools while also considering the broader context of racial disparities in the criminal justice system.

4.4. Training and Development

Getting practitioners and all stakeholders involved in the drive for fairness by giving them the necessary information and resources is pivotal to having a fair ADS system. Providing training and education to employees about fairness, bias, and diversity can help foster a culture of fairness within an organization. Unfavourable outcomes in ADS can sometimes result from oversight by developers or clients. It is crucial to make stakeholders aware of potential risks and how to mitigate them through early training (Curto *et al.,* 2023; Cramer *et al.,* 2018). Equipping employees with the knowledge and skills to recognize and address biases in algorithmic systems can contribute to more equitable decision-making processes. Organizations need to recognize the importance of training and developing practitioners in various essential skills, including knowledge management (such as data training and emotional intelligence), communication, accountability, and transparency. Despite their distinctiveness, these terms are sometimes used interchangeably, collectively referred to as "explicability." Investing in training programs that encompass these skills is essential for ensuring that practitioners possess the necessary competencies to navigate the complex landscape of algorithmic decision-making effectively.

Ferrara *et al.,* (2023); Ueda *et al.,* (2023); Xivuri *et al.,* (2023). Enhancing knowledge management equips practitioners with the ability to comprehend and leverage data effectively, while also fostering emotional intelligence to navigate the human aspects of decision-making processes. Effective communication serves as a cornerstone for conveying complex concepts and system performance to stakeholders. Accountability ensures that practitioners take ownership of their decisions and actions, fostering trust and reliability in the ADS ecosystem. Transparency, meanwhile, entails openness and clarity in disclosing information related to algorithmic processes, cultivating trust and understanding among stakeholders.

Adensamer *et al.,* (2021); Aysolmaz *et al.,* (2023); Kochling *et al,.* (2020); Lee (2018); Lepri *et al.,* (2017); Metcalf *et al.,* (2021); Pessach (2022); Raji *et al.,* (2020); Rana *et al.,* (2023); Ueda *et al.,* (2023); Wang *et al.,* (2020); Veale *et al.,* (2018); Xivuri *et al.,* (2023). Emphasizing these skills and providing comprehensive training programs empowers practitioners to uphold ethical standards and champion fairness in algorithmic decision-making processes. Communication's pivotal role in ensuring effective system performance cannot be overstated. Designers sometimes encounter challenges in accurately communicating the system's performance to users, underscoring the significance of internal communication. Given the variability of decision support across domains, tailored communication strategies are indispensable for effectively conveying model limitations. Lack of communication and understanding can impede fair outcomes, particularly in interactions between decision subjects and administrative systems. Hermann (2022) emphasizes the importance of communication, explainability, accountability, and transparency—concepts he collectively refers to as "explicability." Despite their controversial and often interchangeable use in research, Hermann's study highlights the significance of each factor in creating a fair ADS. Ensuring explicability is crucial for fostering trust and ensuring the ethical and effective implementation of fair ADS.

Transparency and explanation in decision-making processes are highly valued by citizens and clients, as clear elucidations enhance accountability and fairness. By prioritizing these skills and offering

comprehensive training initiatives, organizations foster a culture of ethical responsibility and promote fairness in algorithmic decision-making processes. Establishing formal channels for communication between teams responsible for data collection and model development is paramount (Ferrara *et al.,* 2023; Veale *et al.,* 2018). This can be accomplished through regular meetings, workshops, or collaborative platforms designed specifically for discussing data-related issues, sharing insights, and coordinating efforts to address fairness concerns(Holstein *et al.,* 2019). By formalizing these channels, organizations ensure that pertinent information flows seamlessly between teams, facilitating a holistic approach to addressing fairness considerations throughout the data collection and model development processes.

Regular meetings provide opportunities for teams to discuss progress, identify challenges, and brainstorm potential solutions collaboratively. Workshops offer structured environments for in-depth discussions and skill-building exercises, enabling team members to deepen their understanding of fairness issues and develop practical strategies for mitigating biases in data and models. Collaborative platforms serve as virtual hubs where team members can exchange ideas, share resources, and document best practices in real-time, fostering ongoing communication and collaboration beyond scheduled meetings and workshops. By implementing formal communication channels, organizations demonstrate their commitment to fostering transparency, accountability, and collaboration in addressing fairness concerns within algorithmic decision-making processes (Kochling *et al.,* 2020).

Transparency is crucial for ensuring accountability and trust in machine learning systems. Veale *et al.* (2018) emphasized the internal pressure to provide more explanation behind the system's design, which leads to a transparent result and better stakeholder buy-in. This transparency in model development, specifically the logic behind system design, is essential for fostering trust and understanding among stakeholders. Metcalf *et al.,* (2021); Lepri *et al.,* (2017) Organizations can enhance transparency and accountability by sharing knowledge of the system with stakeholders, including media organizations and journalists. However, transparency may lead to external actors attempting to manipulate systems, posing risks such as gaming or strategic withholding of consent. Therefore, organizations must exercise discretion even while being transparent to prevent unintended consequences and biases. Fairness considerations also impact model interpretability, especially in complex neural networks. Practitioners often prefer simpler, more interpretable models like rule-based decision trees to ensure fairness and maintain transparency (Ferrara *et al.,* 2023). These models facilitate understanding and scrutiny of the decision-making process, thus promoting fairness in ADS systems' deployment and decision-making processes within public sector organizations.

Table 7: Accountability

Accountability to Decision Subject	Transparency
	Detailed Explanation (Communication)
	Knowledge sharing

4.5. Ensuring Fairness in Algorithmic Decision-support: A case Study in Law Enforcement

1. Single Threshold Rule (Threshold Setting): Setting thresholds for detaining individuals based on their likelihood of committing a violent crime can have significant implications for fairness. To mitigate the risk of unfairness, several strategies can be employed. *Collecting more comprehensive and accurate data can improve the estimation of risk***,** thereby lowering the error rate in predictions. Additionally, increasing the threshold for detaining individuals can reduce the number of erroneous detentions across all racial groups, promoting fairness. Furthermore, modifying the decision-making process to minimize the impact of classification errors can help ensure that the consequences of errors are less severe. By implementing these strategies, the fairness and accuracy of detention decisions can be improved, reducing the potential for bias and unjust outcomes (Corbett-Davies *et al.,* 2017).

2. Type of Bias- Historical Bias: When training data, it is crucial to be aware of the types of biases that may be involved in data collection to mitigate unfairness in the training process. The accuracy and bias in data used to train algorithms can significantly impact fairness (Green 2022). For example, biased data, such as higher arrest rates for certain racial groups due to policing practices, can lead to unfair outcomes if not properly accounted for. By recognizing and addressing these biases during data training, we can work towards more equitable and just algorithmic outcomes.

3. Considerations for Feature Selection and Correlation Analysis: *The decision to include or exclude certain features in algorithms can significantly influence fairness*, making careful consideration of potential impacts essential for ensuring equitable outcomes. Balancing statistical robustness with legal and ethical considerations is crucial in designing fair algorithms. During data collection, it is important to check for correlations between variables to avoid reinforcing biases and to enhance the overall fairness of the algorithm (Srinivasan and Chander 2021).

4. Group vs. Individual Choices: Some decisions are better conceptualised as group choices rather than purely individual ones (Lepri *et al.,* 2023). The *distinction between group and individual choices is a crucial factor in determining fairness*, especially when decisions have implications beyond the individual level and aim to promote diversity, equity, and inclusion within communities or institutions.

4.6. Discussion

The introduction, adoption, and implementation of algorithmic decision support (ADS) systems have raised significant concerns about fairness. This issue has attracted numerous scholars to explore various approaches to address it. While some researchers suggest addressing fairness from a socio-technical perspective, others have developed fairness toolkits, such as checklists and auditing tools. Despite these efforts, there remains a need to understand how organizational frameworks can influence fairness without necessitating a complete overhaul of organizational culture and principles. This issue has sparked considerable interest among researchers, particularly regarding how to accommodate practitioners' needs. To understand different recommendations for identifying

organizational factors that can be integrated into managerial practices to promote fairness, a systematic literature review of twenty eight (28) articles was conducted using a modified version of the Wieranga's (2020) protocol. The research aims to identify organizational factors that influence fairness in the context of ADS.

While there is a significant body of literature attempting to address fairness in algorithmic decision support systems, there remains a gap in empirical research on how these concepts manifest in real-life contexts. Based on the review of articles, 26 factors were identified and classified based on implementation – including organizational factors (related to the organization as a whole, its culture, policies, structure, and processes), leadership-related factors (style and characteristics of leaders), and task-based factors (related to team composition and task execution) – as well as thematic factors (governance, social responsibility, technical aspects, and training & development).

The findings underscore the pivotal role organizations play in designing, developing, and implementing fair ADS systems. From pre-processing to post-processing stages, organizations must proactively address fairness concerns. Achieving fairness requires adopting both thematic and implementation-based approaches in ADS processes. Organizations must integrate fairness goals, foster transparent communication among stakeholders, employ clear and transparent models, and encourage collaboration across diverse teams. Providing practitioners with necessary support and training empowers them to identify and address issues effectively.

4.7. Practical Implications

4.71. Governance

Organizations need to recognize and mitigate economic disparities that may hinder equitable access to technological advancements and best practices (Veale *et al.,* 2018). By bridging the gap between larger, wealthier organizations and smaller, resource-constrained entities, collaborative efforts can foster innovation and promote fairness across the ecosystem. Organization need to prioritize fairness and accountability, cultivating a culture of fairness that underpins sustainable success and societal impact. Rather than reacting to fairness issues as they arise, teams need to implement proactive auditing processes to identify and mitigate potential biases before deployment as biases manifest at various stages of the development pipeline, such as the data collection, algorithm design, and decision-making processes. When unfairness is detected, teams need clear guidelines and decision-making frameworks to determine the appropriate course of action to address these instances.

Developing clear and meaningful performance metrics that align with stakeholders' needs and domain expertise is essential for fostering fair ADS systems. Managing stakeholders' expectations and conducting regular impact assessments are integral components of fostering a socially responsible organizational culture. Organizations should regularly evaluate the fairness of their algorithmic systems and be willing to make adjustments as needed. This includes conducting audits, soliciting feedback from stakeholders, and staying informed about developments in fairness research and best practices.

Some organizations may lack the necessary technological infrastructure for the optimal implementation of ADS systems, highlighting the importance of resource provision. Organizations

should be prepared to recruit experts when necessary to facilitate the implementation process and seek clarification as needed to comprehend system functionality thoroughly. Adequate resource allocation from both organizations and policymakers is crucial to ensure effective system utilization.

4.72. Social Responsibility

Collaboration is another critical factor frequently mentioned in the reviewed articles. Organizations should foster collaboration across teams, involving stakeholders from diverse backgrounds and interdisciplinary teams. Having people with varied perspectives contributes to a more robust and fair system design. Organizations are need to assemble multidisciplinary research teams with diverse perspectives. By integrating insights from fields such as ethics, sociology, computer science, and law, these teams can foster dynamic collaborations and innovative solutions. This interdisciplinary approach enriches the design process and enhances the likelihood of achieving a fairer Algorithmic Decision-Support System.

Human intervention often becomes necessary to enrich algorithmic outputs with contextual data, ensuring decisions are informed by both data-driven insights and institutional knowledge. Designing efficient interfaces for human-machine collaboration is imperative to enhance decision-making processes. Additionally, the social practices associated with algorithmic systems significantly influence their effectiveness and ethical implications. Documenting and transferring these practices across various contexts is crucial to ensure consistency and mitigate biases.

Organizations need to enable users to make context-based decisions and create flexible options within decision-support systems, fostering a human-centric approach to ADS design and implementation. For fair and effective implementation, it's crucial to consider broader socio-economic contexts and the specific needs of organizations when sharing models. Human intervention is often necessary to complement algorithmic outputs with contextual data, ensuring decisions are informed by both data-driven insights and institutional knowledge. Designing effective interfaces for human-machine collaboration is vital for improving decision-making processes. The social practices surrounding algorithmic systems significantly impact their effectiveness and ethical implications. Documenting and transferring these practices across contexts is essential for ensuring consistency and mitigating biases.

4.73. Technical

Design and application of algorithms play a pivotal role in determining fairness, as they have the potential to enhance both efficiency and equity. However, implementing these systems presents complex challenges that demand careful scrutiny from researchers and policymakers. The findings highlight the significant impact of data management on ADS fairness, elucidating on how organizations collect their data plays a crucial role; the data must be of high quality, representative of key features, and accurate. When necessary, expanding data collection can provide a more comprehensive and precise dataset.

Organizations must proactively support practitioners in collecting and curating high-quality datasets to ensure fairness throughout the lifecycle of algorithmic decision-making. Emphasis on quality dataset for a fair system (Cramer *et al.,* 2018). The necessity for novel methodologies, processes,

metrics, and tools to handle fairness has been identified, underscoring the importance of enhancing data management practices to mitigate biases and adhere to specific fairness constraints. This fosters equity and transparency in algorithmic decision-making processes. Developing tools and methodologies to guide data collection and curation processes is essential, as teams often struggle to identify relevant subpopulations for data collection, leading to imbalanced datasets. Techniques for identifying and mitigating biases during data collection, along with strategies to ensure diversity and representativeness in datasets, are crucial components.

4.74. Training and Development

Organizations need to allocate resources towards continuous training and education initiatives. These programs should aim to increase awareness of bias and fairness issues among developers, implementers, and users of algorithmic systems. Empowering practitioners to identify and address errors throughout every stage of development to implementation is also essential. A culture of accountability and transparency must be cultivated within the organization. This involves fostering open communication between different units internally and with clients externally. By ensuring that all stakeholders are informed and involved in the decision-making process, organizations can enhance trust and mitigate the risk of bias. When designing algorithms, developers should prioritize incorporating human oversight mechanisms. This ensures that decisions made by the system are subject to human review and intervention when necessary. By combining technological advancements with human judgment, organizations can uphold fairness and equity in their decision-making processes.

Organizations utilizing algorithmic decision support systems (ADS) must prioritize transparency, accountability, and user empowerment. Users should have comprehensive access to information regarding algorithmic operations, including data inputs, decision-making procedures, and potential biases within the system.

Overall, the culture and values within an organization play a significant role in shaping fairness. Organizational cultures that prioritize diversity, equity, and inclusion are more likely to foster fair decision-making processes. Conversely, cultures that tolerate or perpetuate discrimination can undermine fairness in algorithmic systems. Furthermore, establishing robust feedback mechanisms is essential for addressing instances of unfairness or bias in decision outcomes. Empowering users to challenge algorithmic decisions and providing avenues for redress can mitigate the adverse effects of biases, fostering fairness and transparency in decision-making processes. Moreso, fostering fairness in algorithmic decision support requires a quality and representative dataset, collaborative efforts, proactive measures, and a commitment to ethical standards and transparency.

4.7. Recommendation and future directions

This paper provides an overview of organizational factors that could influence fairness. However, some aspects are beyond the scope of this study, such as factors influencing fairness in specific sectors like healthcare and finance. One key factor influencing fairness is the consideration of context-specificity when designing or implementing ADS systems. Since system design is not a one-size-fits-all solution, it is crucial to consider the context and ensure proper representation of data.

Future research could investigate factors that influence fairness from a sector-specific perspective rather than a general one.

4.71. Recommended Approaches for ensuring fairness in ADS

Studies such as Ferrara et al. (2023), Makhlouf et al. (2021), and Carey and Wu (2022) have suggested new methodological approaches to addressing fairness in organizations. Building on their research, this study proposes the following approaches to enhance fairness in organizations:

Statistical Notions of Fairness: Statistical notions of fairness provide a framework for addressing biases and discrimination in algorithmic decision-making. Practitioners often rely on statistical tests such as the chi-squared test or t-test to assess whether algorithmic outcomes are biased across different demographic groups. These tests help analyze the impact of algorithms on diverse populations and monitor fairness. By scrutinizing the impact of algorithms on various demographic groups, practitioners can work towards equitable outcomes in decision-making processes. However, the application of statistical notions can vary depending on the specific context and domain, so practitioners must carefully consider which tests are appropriate for their use case

Similarity-Based Notions of Fairness: Practitioners use similarity functions to identify correlations among similar individuals and ensure non-discriminatory predictions. However, determining relevant attributes and measuring similarity can be complex and context-dependent. Despite these challenges, similarity-based notions are widely used to manage fairness in various contexts, such as education and job hiring.

Causal Notions of Fairness: Causal inference approaches are employed to identify and mitigate sources of unfairness by examining the causal relationships between variables. Understanding these relationships helps identify mechanisms through which biases are introduced, leading to the development of fairer scoring methodologies. However, implementing causal notions of fairness can be complex due to the difficulty of computing causal graphs.

Substantive Algorithmic Fairness: Substantive algorithmic fairness represents a departure from the traditional approach of formal algorithmic fairness, which relies on mathematical models to address discrimination and bias in algorithms. Instead, it proposes integrating social justice principles into algorithmic decision-making. This approach emphasizes that fairness in algorithms goes beyond merely avoiding discrimination or bias in their outcomes; it aims to promote justice, equity, and societal well-being. Substantive algorithmic fairness seeks to ensure that algorithms contribute to a more just and equitable society by emphasizing fairness, accountability, and ethical responsibility in their design, implementation, and use (Green 2022).

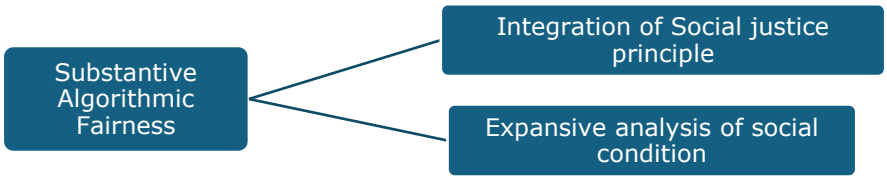

Figure 7: Substantive Algorithm

Expansive Analysis of Social Conditions: Substantive algorithmic fairness involves analyzing social conditions and institutions to understand the broader context in which algorithmic decisions are made. This includes considering relational and structural factors that influence decision points. The term "socio-contextual analysis" could aptly describe the process of conducting an expansive analysis of social conditions as it relates to algorithmic decision-making. This terminomogies emphasizes the examination of the broader social context surrounding algorithmic systems, including factors such as historical inequalities, structural injustices, and relational dynamics. It conveys the idea of going beyond purely technical considerations to understand the social complexities that shape algorithmic outcomes and implications. Additionally, "contextual evaluation" or "socio-structural assessment" could also be used to convey a similar concept of analyzing the broader social conditions influencing algorithmic fairness.

Integration of Social Justice Principles: Rather than relying solely on formal mathematical models of fairness, substantive algorithmic fairness integrates principles of social justice into algorithmic decision-making. It aims to combat social hierarchies and promote equitable public policy. The integration of social justice principles into algorithmic decision-making could be referred to as "justice-aligned algorithm design" or "equity-centered algorithm development." These terms highlight the deliberate effort to incorporate principles of social justice, fairness, and equity into the design, deployment, and evaluation of algorithms. Additionally, "ethical algorithmic engineering" or "fairness-conscious algorithm design" could also convey the notion of embedding social justice principles into algorithmic processes. These terms emphasize the ethical considerations and fairness criteria that guide the development of algorithms with a focus on promoting equitable outcomes and addressing societal inequalities.

By adopting these approaches, organizations can improve the fairness of their ADS and align them with broader socio-contextual needs.

Based on various articles reviewed, the following actions and recommendations can be proposed:

1. Invest in Training: Organizations should invest in training stakeholders on the importance of transparency and explainability in design, encouraging each stakeholder to take accountability for fairness. Moreso, organization need to invest in training and providing resources for data collection, curation, and model evaluation is vital for fostering a culture of fairness and accountability in algorithmic decision-making processes.

2. Develop Novel Methodologies: It is essential to develop new methodologies and processes for model development.

3. Understand Social Contexts: Understanding the interplay between algorithmic decision-making and social hierarchies will help organizations develop fair systems. Developing frameworks for balancing competing societal values in algorithmic decision-making processes is crucial.

4. Interdisciplinary Collaboration: Researchers and practitioners need to collaborate closely to achieve fair outcomes. More research is required to develop tools for auditing, essential fairness metrics, and integrating ethics policies into the core organizational system to ensure the successful and fair adoption of ADS (Heaton *et al.,* 2023).

6. Facilitate Knowledge Sharing: Promote cross-team knowledge sharing among diverse backgrounds to identify blind spots and deepen the understanding of fairness issues.

By implementing these actions and recommendations, organizations can foster a culture of ethical responsibility and promote fairness in algorithmic decision-making processes.

4.72. Research gaps and future work

This field is rapidly evolving, necessitating more in-depth research into the factors that could enhance fairness. Currently, there is a lack of sufficient empirical research on the factors influencing fairness in organizations, with most available articles focusing on accountability. While accountability is crucial for fair algorithmic decision-support design, more empirical research is needed to address fairness factors beyond accountability. There is a need for more research on practical contexts to identify factors that could influence fairness.

To contribute to this, some researchers have identified the need for fairness metrics and auditing tools. Although the literature indicates that several toolkits are already available in the market, researchers like Lee and Singh (2021) have compared different fairness metrics, highlighting the importance of selecting appropriate tools for specific conditions. Despite the availability of these tools, further research is needed to establish explicit conditions for selecting fairness metrics and to develop equitable methods for implementing them. Similar suggestions were noted by Madaio et al. (2020), underscoring the importance of this issue to the body of knowledge.

Furthermore, researchers could explore the correlation between specific contexts and feature selection impacts, as well as trade-offs in influencing fairness. Future research could examine the relationship between ethical considerations and the context specificity of fairness, as well as how substantive algorithmic fairness can be operationalized in real-world contexts.

There has been a notable discrepancy in research regarding the incorporation of human judgment into automated decision-making systems (ADS). While human judgment holds the potential to introduce bias, it also presents an opportunity for fair collaboration between humans and ADS. Investigating tools that facilitate equitable cooperation between human judgment and ADS is another area of interest that warrants further exploration.

The findings of this study could be further enhanced by employing alternative research approaches, such as qualitative methods like interviews, questionnaires, or case studies, to identify organizational factors that influence fairness. Given that this is a relatively new area of research and not yet fully explored, there were fewer related articles available for review.

4.8. Limitations

The study focused on identifying organizational factors influencing fairness in algorithmic decision support (ADS). This area proved challenging to explore, as most studies concentrate on mitigating bias in machine learning (ML) or artificial intelligence (AI). For this research, it was assumed that ML, AI, and ADS are interrelated, with ADS considered a subset of ML and AI. This assumption allowed for a broader inclusion of articles for review, though the exact correlations between these

terms were not explicitly addressed. Additionally, factors mitigating bias were treated as analogous to factors influencing fairness, potentially affecting the results.

The selection of search strings was limited to English-language studies, which means the review cannot claim to be exhaustive. For future research, incorporating related terms and including non-English studies could provide a more comprehensive understanding. Despite these limitations, several factors were identified that could support organizations in developing and implementing fair ADS systems.

Future research should explicitly differentiate between ML, AI, and ADS to clarify their relationships and impacts. Additionally, considering a wider array of search terms and including diverse languages will help ensure a more thorough exploration of organizational factors influencing fairness.

5.0 CONCLUSION

This thesis reviews literature on the organizational factors that influence fairness in algorithmic decision support systems (ADS), explores organizational factors influencing ADS, considering both the perspective of organizations developing these systems and the organizations who adopt their systems for their use. This research contributed to the body of knowledge by categorizing these factors into two angles in which these factors can be assessed; Theme based (including organizational cultures, policies, ethical guidelines etc.) and Implementation based (such as team composition -task to be carried out, stakeholders involvement). Several key factors that shape fairness were identified in this review: organizational culture, leadership commitment, auditing practices, integration of fairness metrics/checklists into organizational goals, cross field collaboration, data management, feature selection, Human judgement and resource allocation. Understanding these factors and ensuring that fairness is prioritized at all levels, from top leadership to frontline users, is essential for fostering a fair decision-making environment.

For organizations employing ADS, it is also crucial for leaders to integrate fairness into their goals, prioritize it, and allocate sufficient resources for effective implementation. For ADS to be truly fair, resources must be adequately allocated, and the organizational infrastructure must be capable of supporting the system. Investing in robust systems and hiring experts are necessary steps for achieving fair and effective results.

The findings of this study highlight the importance of a stakeholder co-creation approach in ensuring fairness. Since fairness concerns are socio-technical issues, collaboration across fields and the creation of checklists and ethical guidelines are vital. This thesis contributes to the discussion on fairness in ADS design by emphasizing the need for collaboration, ethical guidelines, and toolkits to provide a transparent, accurate, and fair system for shareholders and stakeholders. Additionally, it underscores the necessity of training for all parties involved and identifies areas where organizations should focus to achieve a fair ADS system, which have not been adequately addressed in existing literature.

In conclusion, fairness considerations extend beyond accuracy and explainability, affecting various aspects of ADS systems. For instance, in healthcare, balancing fairness with user privacy is paramount, as sharing sensitive medical data for training models raises privacy concerns despite potentially improving fairness. Similarly, in bioinformatics, systematic bias in high-throughput processing can impact model efficiency, highlighting the need to address fairness issues comprehensively across different domains. Achieving fairness in algorithmic decision support requires a multifaceted approach that acknowledges the varied perspectives of deploying organizations and system users, considering the specific context and domain in which these systems operate. Fairness perception is deeply influenced by these nuances. By addressing key organizational factors and championing transparency, accountability, and clear communication, organizations can cultivate an environment conducive to the fair development, deployment, and implementation of ADS. This approach not only mitigates biases but also maximizes the advantages of ADS, ensuring their effectiveness and integrity in diverse operational settings.

REFERENCES

Adensamer, A., Gsenger, R., & Klausne, L. D. (2021). "Computer says no": Algorithmic decision support and organisational responsibility. *Journal of Responsible Technology*, *Volumes 7–8,*, Pg. 1 - 10. Sciencedirect. https://doi.org/10.1016/j.jrt.2021.100014

AlgorithmWatch. (2019). Automating society: Taking stock of automated decision-making in the EU. Berlin. Available at: https://algorithmwatch.org/wpcontent/uploads/2019/01/Automating_Society_Report_2019.pdf (accessed 18 June 2024).

Alkhatib, A., & Bernstein, M. S. (2019). Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-13). Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973-989.

Araujo, T., Helberger, N., Kruikemeier, S., & De Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society*, *35*(3), 611–623. https://doi.org/10.1007/s00146-019-00931-w

Arrow, K. J. (1973). The Theory of Discrimination, S. 3-33 in: Orley Ashenfelter und Albert Rees (Hg.): Discrimination in Labor Markets. *The Theory of Discrimination*, S, 3-33.

Aysolmaz, B., Müller, R., & Meacham, D. (2023). The public perceptions of algorithmic decision-making systems: Results from a large-scale survey. Telematics and Informatics, 79, 101954. https://doi.org/10.1016/j.tele.2023.101954Bannister, F., & Connolly, R. (2011). The trouble with transparency: A critical review of openness in e-government. *Policy & Internet*, 3(1), 1-30.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104, 671.

Belle, A., & Papantonis, I. (2021). Towards explainable artificial intelligence: Four ways to increase explainability. *Expert Systems with Applications*, 176, 114796.

Bovens, M. (2006). Analysing and assessing public accountability. A conceptual framework. European Governance Papers (EUROGOV) No. *No. C-06*, *12006*.

Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework 1. European law journal, 13(4), 447-468.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8).

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. 77–91 pages.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 2053951715622512.

Calders, T., & Verwer, S. (2010). Three naive bayes approaches for discrimination-free classification. Data mining and knowledge discovery, 21, 277-292.

Calders, T., & Žliobaitė, I. (2013). Why unbiased computational processes can lead to discriminative decision procedures. In Discrimination and Privacy in the Information Society: Data mining and profiling in large databases (pp. 43-57). Berlin, Heidelberg: Springer Berlin Heidelberg.

Carey, D., & Smith, M. (2016). How companies are using simulations, competitions, and analytics to hire. *Harvard Business Review*, 22.

Carey, A. N., & Wu, X. (2022). The Causal Fairness Field Guide: Perspectives From Social and Formal Sciences. *Frontiers in Big Data*, *5*. https://doi.org/10.3389/fdata.2022.892837

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015, August). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1721-1730).

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809-825.

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., & Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. *American Economic Review*, 106(5), 124-127.

Chander, A. (2016). The racist algorithm. *Mich. L. Rev.*, 115, 1023.

Chiappa, S. (2019, July). Path-specific counterfactual fairness. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 7801-7808).

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, *5*(2), 153-163.

Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. Communications of the ACM, 63(5), 82-89.

Christin, A. (2017). Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society*, 4(2), 2053951717718855.

Christin, A., Rosenblat, A., & Boyd, D. (2015). Courts and predictive algorithms. Data & Civil Rights.

Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Washington Law Review*, 89, 1-33.

CJEU (Court of Justice of the European Union). (2018). Judgment in Case C-210/16 Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein v Wirtschaftsakademie Schleswig-Holstein GmbH.

CJEU (Court of Justice of the European Union). (2019). Judgment in Case C-40/17 Fashion ID GmbH & Co. KG v Verbraucherzentrale NRW eV.

Cobbe, J., & Singh, J. (2021). Artificial intelligence as a service: Legal responsibilities, liabilities, and policy challenges. *Computer Law & Security Review*, *42*, 105573.

Colander, D. (2022). Toward Understanding the Human Dimensions of Explainable AI. *Journal of Business Ethics*, 1-15.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic Decision Making and the Cost of Fairness. *KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Pg. 797–806. ACM. 10.1145/3097983.3098095

Coston, A., Mishler, A., Kennedy, E. H., & Chouldechova, A. (2020, January). Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 582-593).

Courtland, R. (2018). Bias detectives: The researchers striving to make algorithms fair. Nature. Retrieved March 15, 2022.

Courtois, C., & Timmermans, J. (2018). Algorithmic transparency in the European General Data Protection Regulation: Scrutinizing the "new" accountability. *Computer Law & Security Review*, 34(2), 398-411.

Cramer, H., Garcia-Gathright, J., Springer, A., & Reddy, S. (2018). Assessing and addressing algorithmic bias in practice. *Interactions*, *25*(6), 58–63. https://doi.org/10.1145/3278156

Crawford, K. (2013). The hidden biases in big data. *Harvard business review*, *1*(4).

Curto, G., & Comim, F. (2023). SAF: Stakeholders' Agreement on Fairness in the Practice of Machine Learning Development. Science and Engineering Ethics, 29(4). https://doi.org/10.1007/s11948-023-00448-y

Danaherm, J. (2016). The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology*, 29(3), 245-268.

Danks, D., & London, A. J. (2017, August). Algorithmic Bias in Autonomous Systems. In *Ijcai* (Vol. 17, No. 2017, pp. 4691-4697).

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 1148-1158).

De Vito, M. A., Artstein, R., Bojar, O., Eidelman, V., Gao, J., Hall, K., ... & Strassel, S. (2017). Towards a Definition of Misinformation. In *Proceedings of the 12th International Workshop on Semantic Evaluation* (pp. 540-546).

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020, April). A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-12).

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. Digital Journalism, 3(3), 398-415.

Diakopoulos, N., & Koliska, M. (2016). Algorithmic transparency in the news media. *Digital Journalism*, 4(7), 809-828.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114.

Dotan, R. (2021). Theory choice, non-epistemic values, and machine learning. Synthese, 198(11), 11081-11101.

Dreyer, S., & Schulz, W. (2019). The General Data Protection Regulation and automated decision-making: Will it deliver. *Bertelsmann Stiftung*.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226).

Edwards, H., & Storkey, A. (2015). Censoring representations with an adversary. arXiv preprint arXiv:1511.05897.

Eiband, M., Zlabinger, M., Aryal, S., Schöning, J., & Löchtefeld, M. (2018). Towards a taxonomy of explainability techniques in algorithmic decision-making systems. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1-13.

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

European Commission. (2020). *White Paper on Artificial Intelligence: A European Approach to Excellence and Trust*. Brussels: European Commission.

European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L 119/1.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015, August). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 259-268).

Ferguson, A. G. (2016). Policing predictive policing. *Wash. UL Rev.*, 94, 1109.

Ferrara, C., Sellitto, G., Ferrucci, F., Palomba, F., & De Lucia, A. (2023, November 24). Fairness-aware machine learning engineering: how far are we? *Empirical Software Engineering*, *29*(9), Pg. 1 - 46. Springer. 10.1007/s10664-023-10402-y

Gendler, T. S. (2011). On the epistemic costs of implicit bias. Philosophical Studies, 156(1), 33-63.

Gillespie, T. (2016). # trendingistrending: When algorithms become culture. In *Algorithmic Cultures* (pp. 64-87). Routledge.

Green Chen, Y. (2019). Explaining How: Structured Explanations of Decisions from Small Data Sets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 4242-4249.

Green, B. (2022, October 08). Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness. Philosophy & Technology, 35(90). Springer. https://doi.org/10.1007/s13347-022-00584-6

Green, B. (2018). The silent normative assumptions of the "normative" in "AI ethics". In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 1-7).

Green, B. (2018b). Unpacking the Invisible Knapsack: The Intersection of Normativity and Bias in AI Ethics. *Journal of Artificial Intelligence Research*, 63, 715-745.

Green, B., & Chen, Y. (2019, January). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 90-99).

Gürses, S., & Van Hoboken, J. (2018). Privacy after the Agile Turn. In E. Selinger, J. Polonetsky, & O. Tene (Eds.), Cambridge Handbook of Consumer Privacy (pp. 579–601), Cambridge University Press.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, *29*.

Hashimoto, T., Srivastava, M., Namkoong, H., & Liang, P. (2018, July). Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning* (pp. 1929-1938). PMLR

Heaton, D., Clos, J., Nichele, E., & Fischer, J. E. (2023). The Social Impact of Decision-Making Algorithms: Reviewing the Influence of Agency, Responsibility and Accountability on Trust and Blame. *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*. https://doi.org/10.1145/3597512.3599706

Hermann, E. (2022). Leveraging Artificial Intelligence in Marketing for Social Good—An Ethical Perspective. *Journal of Business Ethics*, *179*(1), 43–61. https://doi.org/10.1007/s10551-021-04843-y

Holstein, K., Vaughan, J. W., Daumé, H., Dudik, M., & Wallach, H. (2019). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, *600*, Pg. 1 - 16. ACM. 10.1145/3290605.3300830

Hunkenschroer, T., & Luetge, C. (2022). Accountability in Artificial Intelligence: Towards Transparency and Interpretability in Machine Learning. *Journal of Business Ethics*, 1-20.

Hutchinson, B., & Mitchell, M. (2019, January). 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 49-58).

Imai, K., & Jiang, Z. (2023). Principal fairness for human and algorithmic decision-making. *Statistical Science*, *38*(2), 317-328.

Johndrow, J. E., & Lum, K. (2019). An algorithm for removing sensitive information. *The Annals of Applied Statistics*, *13*(1), 189-220.

Johnson, K. N. (2019). Creating Just and Moral Corporate Structures: A Necessity for Algorithmic Fairness Oversight. *Journal of Business Ethics*, 1-17.

Joseph, M., Kearns, M., Morgenstern, J., Neel, S., & Roth, A. (2016). Rawlsian fairness for machine learning. *arXiv preprint arXiv:1610.09559*, *1*(2), 19.

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, *33*(1), 1-33.

Kamishima, T., Akaho, S., & Sakuma, J. (2011, December). Fairness-aware learning through regularization approach. In *2011 IEEE 11th international conference on data mining workshops* (pp. 643-650). IEEE.

Kay, M., Matuszek, C., & Munson, S. A. (2015, April). Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems* (pp. 3819-3828).

Kearns, M., & Roth, A. (2019). The ethical algorithm: The science of socially aware algorithm design. Oxford University Press.

Kern, C., Gerdon, F., Bach, R. L., Keusch, F., & Kreuter, F. (2022, October 14). Humans versus machines: Who is perceived to decide fairer? Experimental evidence on attitudes toward automated decision-making. *Patterns*, *Vol 3*(10). PubMed. 10.1016/j.patter.2022.100591

Kim, P. T. (2016). Data-driven discrimination at work. *Wm. & Mary L. Rev.*, 58, 857.

Kirkpatrick, K. (2016). Battling algorithmic bias. *Communications of the ACM*, *59*(10), 16–17. https://doi.org/10.1145/2983270

Kleanthous, S., Kasinidou, M., Barlas, P., & Otterbacher, J. (2022, January 14). Perception of fairness in algorithmic decisions: Future developers' perspective. *Patterns*, *Vol. 3*(1), Pg. 1 -18. Science Direct. https://doi.org/10.1016/j.patter.2021.100380

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, *133*(1), 237-293.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference* (pp. 43:1-43:23).

Köchling, A., & Wehner, M. C. (2020, November 20). *Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development*, *13*, Pg. 795 - 848. Springer. https://doi.org/10.1007/s40685-020-00134-w

Koefer, F., Lemken, I., & Pauls, J. (2023, April). Fairness in Algorithmic Decision Systems: A Microfinance Perspective. European Investment Fund (EIF), Pg. 1 - 32. https://www.findevgateway.org/paper/2023/04/fairness-algorithmic-decision-systems-microfinance-perspective

Koutsouris, F. A. C. (n.d.). *When the algorithm is your teacher: Perceptions of fairness, trust, and decision*.

Kraemer, S., van Overveld, K., & Peterson, M. (2011). Responsibility for the unintended consequences of actions. In *Ethics and the Internet in West Africa: Toward an Ethical Model of Integration* (pp. 55-72). IGI Global.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in neural information processing systems*, *30*.

Langer, M., König, C. J., & Papathanasiou, M. (2019). Highly automated job interviews: Acceptance under the influence of stakes. *International Journal of Selection and Assessment*, 27(3), 217-234.

Lee, J. M., Son, H. Y., Kang, H., Jung, S., & Kim, J. H. (2015). Investigating the impacts of transparency and interaction on mobile application. *International Journal of Human-Computer Interaction*, 31(8), 523-532.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1), 2053951718756684.

Lee, N. T. (2018). Policies for Countering Algorithmic Bias: Incorporating Cultural Diversity into Data. *Journal of Business Ethics*, 1-20.

Lee, M. S. A., & Singh, J. (2021). The Landscape and Gaps in Open Source Fairness Toolkits. https://doi.org/10.1145/3411764.3445261

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. Philosophy & Technology, 31(4), 611-627.

Leventhal, G. S. (1980). What Should Be Done With Equity Theory? In *Social Exchange: Advances in Theory and Research* (pp. 27-55). Springer.

Lindebaum, D., Vesa, M., & Den Hond, F. (2020). Insights from "The Machine Stops" to better understand rational assumptions in algorithmic decision making and its implications for organizations. *Academy of Management Review*, 45(1), 247-263.

Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2012). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 623-631).

MacKenzie, D. (2014). The Sociology of Algorithms: High-Frequency Trading and the Shaping of Markets. *Social Studies of Science*, 44(3), 466-489.

Madaio, M. A., Stark, L., Vaughan, J. W., & Wallach, H. (2020). Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1 -14. ACM. 10.1145/3313831.3376445

Madaio, M., Egede, L., Subramonyam, H., Vaughan, J. W., & Wallach, H. (2022). Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. Proceedings of the ACM on Human-computer Interaction, 6(CSCW1), 1–26. https://doi.org/10.1145/3512899

Mahieu, R., Van Hoboken, J., & Asghari, H. (2019). Responsibility for Data Protection in a Networked World: On the Question of the Controller, Effective and Complete Protection and Its Application to Data Access Rights in Europe. *J. Intell. Prop. Info. Tech. & Elec. Com. L.*, *10*, 84.

Makhlouf, K., Zhioua, S., & Palamidessi, C. (2021). Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management*, *58*(5), 102642. https://doi.org/10.1016/j.ipm.2021.102642

Marabelli, M., Newell, S., Handunge, V. (2021). The lifecycle of algorithmic decision-making systems: Organizational choices and ethical challenges. Journal of Strategic Information System. Vol. 30. Elsevier. https://www.sciencedirect.com/science/article/pii/S0963868721000305

Marcinkowski, F., Kieslich, K., Starke, C., & Lünich, M. (2020). *Implications of AI (un-)fairness in higher education admissions*. https://doi.org/10.1145/3351095.3372867

Martin, K. (2018, June 07). Ethical Implications and Accountability of Algorithms. Journal of Business Ethics, 160, Pg. 835 - 850. Springer Link. https://doi.org/10.1007/s10551-018-3921-3

Martin, K. E. (2018). Algorithms, Bias, and Competitive Advantage. *Bus. Law. Rev.*, 72, 27.

Mayson, S. G. (2018). Bias in, bias out. Yale Law Journal, 128, 2218.

McDonald, K., Fisher, S., & Connelly, C. E. (2017). E-HRM systems in support of "smart" workforce management: An exploratory case study of system success. In *Electronic HRM in the Smart Era* (pp. 87-108). Emerald Publishing Limited.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, *54*(6). https://doi.org/10.1145/3457607

Meijer, A. (2014). Understanding modern transparency. *International Review of Administrative Sciences*, 80(2), 205-224.

Metcalf, J., Moss, E., Watkins, E. A., Singh, R., & Elish, M. C. (2021, March 01). Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Pg. 735–746. ACM. https://doi.org/10.1145/3442188.3445935

Millard, C. (2021). AI and data protection law. In *Research Handbook on the Law of Artificial Intelligence* (pp. 290-312). Edward Elgar Publishing.

Mittelstadt, B. (2016). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11), 501.

Möhlmann, M., & Zalmanson, L. (2017). Hands on the wheel: Navigating algorithmic management and Uber drivers' autonomy. In *Proceedings of the International Conference on Information Systems (ICIS)*, Seoul, South Korea (pp. 10-13).

Moritz, P., Nishihara, R., & Jordan, M. (2016, May). A linearly-convergent stochastic L-BFGS algorithm. In *Artificial Intelligence and Statistics* (pp. 249-258). PMLR.

Moses, L. B., & Chan, J. (2018). Algorithmic prediction in policing: Assumptions, evaluation, and accountability. *Policing and Society*, 28, 806.

Nabi, R., & Shpitser, I. (2018, April). Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).

Noriega, M. (2020). The application of artificial intelligence in police interrogations: An analysis addressing the proposed effect AI has on racial and gender bias, cooperation, and false confessions. Futures, 117, 102510.

O'Neil, C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Park, H., Ahn, D., Hosanagar, K., & Lee, J. (2022, April 28). Designing Fair AI in Human Resource Management: Understanding Tensions Surrounding Algorithmic Evaluation and Envisioning Stakeholder-Centered Solutions. *CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, *55*, Pg. 1–22. ACM. https://doi.org/10.1145/3491102.3517672

Pessach, D. (2022, February 03). A Review on Fairness in Machine Learning. *ACM Computing Surveys*, *Vol. 55*(3), Pg. 1-44. ACM Digital Library. https://doi.org/10.1145/3494672

Rader, E. (2017). Beyond accuracy: Understanding the role of user perceptions of algorithmic systems. Journal of the Association for Information Science and Technology, 68(12), 2724-2736.

Rader, E., & Gray, J. (2015). Understanding User Perceptions of Inaccuracy and Bias in Algorithmic Filtering. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 3737-3746).

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 469-481).

Rai, R. (2020). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. Data Science and Analytics, 29.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020, January 27). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Pg. 33 -44. ACM. https://doi.org/10.48550/arXiv.2001.00973

Rambachan, A., & Roth, J. (2019). Bias in, bias out? Evaluating the folk wisdom. arXiv preprint arXiv:1909.08518.

Rana, S. A., Azizul, Z. H., & Awan, A. A. (2023). A step toward building a unified framework for managing AI bias. PeerJ Computer Science, 9, e1630. https://doi.org/10.7717/peerj-cs.1630Rawls, J. (1971). A theory of justice. Cambridge (Mass.).

Renijith, V., Thomas, S., & Mohan, S. (2020). Artificial Intelligence and Machine Learning: A Review. International Journal of Engineering Research & Technology, 9(3), 55-63.

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

Richardson, R., Schultz, J. M., & Southerland, V. M. (2019). Litigating algorithms: 2019 US report. AI Now Institute.

Salih, A., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Menegaz, G., & Lekadir, K. (2023, May 3). *Commentary on explainable artificial intelligence methods: SHAP and LIME*. arXiv.org. https://arxiv.org/abs/2305.02012

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. R. (2019). Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer Nature.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and Discrimination: Converting Critical Concerns into Productive Inquiry.

Savage, D. D., & Bales, R. (2016). Video games in job interviews: Using algorithms to minimize discrimination and unconscious bias. ABAJ Lab. & Emp. L., 32, 211.

Schumann, C., Foster, J. S., Mattei, N., & Dickerson, J. P. (2020). We Need Fairness and Explainability in Algorithmic Hiring. 1716–1720.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59–68. https://doi.org/10.1145/3287560.3287598

Shneiderman, B. (2021, July 26). SHARE ON Responsible AI: bridging from ethics to practice. Communication of the ACM, Vol. 64(8), Pg. 32 - 35. ACM. https://doi.org/10.1145/3445973

Shin, J., & Choi, B. (2014). Socio-Technical Analysis of Algorithmic Fairness: Towards Human-Centered Algorithm Use. Journal of Business Ethics, 1-15.

Shin, J., & Park, Y. J. (2019). Algorithmic transparency in the GDPR: A regulatory perspective. International Data Privacy Law, 9(3), 191-209.

Shin, J., Euiyoung, K., & Dongwook, H. (2020). Algorithmic transparency in e-commerce. Telematics and Informatics, 52, 101416.

Silverman, R. E., & Waller, N. (2015). The algorithm that tells the boss who might quit. Wall Street Journal.

Simbeck, K. (2019). HR analytics and ethics. IBM Journal of Research and Development, 63(4/5), 9-1.

Simkute, I., Luger, E., & Evans Jones, R. (2020). How Do Users Really Interact with Machine Learning Systems? An Exploratory Study of User Experiences and Design Challenges. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1-14.

Skarlicki, D. P., & Folger, R. (1997). Retaliation in the workplace: The roles of distributive, procedural, and interactional justice. Journal of Applied Psychology, 82(3), 434.

Sloan, L., & Warner, W. (2017). The Science of Fake News. Science, 359(6380), 1094-1096

Srinivasan, R., & Chander, A. (2021). Biases in AI Systems. ACM Queue, 19(2), 45–64. https://doi.org/10.1145/3466132.3466134

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2022, October 10). Fairness Perceptions of algorithmic decision-making: A sytematic review of the empirical literature. Big Data & Society, Vol. 9(2). Sage. https://doi.org/10.1177/20539517221115189

Suen, H. Y., Chen, M. Y. C., & Lu, S. H. (2019). Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes? Computers in Human Behavior, 98, 93-101.

Sullivan, B., & Fosso Wamba, S. (2022). Prioritizing Social and Ethical Considerations in Building Fair AI: The Role of AI Developers, Organizations, and Policymakers. Journal of Business Ethics, 1-19.

Sundar, S. S., & Nass, C. (2001). Conceptualizing sources in online news. Journal of Communication, 51(1), 52-72.

Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. arXiv preprint arXiv:1901.10002, 2(8), 73.

Torresen, J. (2018). Ethical Considerations and Challenges in AI: Towards an Ethical Code of Conduct. In Artificial Intelligence in Education (pp. 3-10). Springer, Cham.

Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R., Matsui, Y., Nozaki, T., Nakaura, T., Fujima, N., Tatsugami, F., Yanagawa, M., Hirata, K., Yamada, A., Tsuboyama, T., Kawamura, M., Fujioka, T., & Naganawa, S. (2023, August 04). Fairness of artificial intelligence in healthcare: review and recommendations. Japanese Journal of Radiology, 42, Pg. 3 -15. Springer. https://doi.org/10.1007/s11604-023-01474-3

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society, 4(2), 20

Veale, M., Kleek, M. V., & Binns, R. (2018). Fairness and Accountability Design Needs for Algorithmic Support in High-stakes Public Sector Decision-Making. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (2018), 440(-), 1-14. ACM. 10.1145/3173574.3174014539517177435 30.

Vedder, A., & Naudts, L. (2017). Accountability for the use of algorithms in a big data environment. International Review of Law, Computers & Technology, 31(2), 206-224.

Vlasceanu, M., Dudik, M., & Momennejad, I. (2022). Interdisciplinarity, Gender Diversity, and Network Structure Predict the Centrality of AI Organizations. 2022 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3531146.3533069

Wang, M., & Deng, W. (2020). Mitigating bias in face recognition using skewness-aware reinforcement learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9322-9331).

Wang, R., Harper, F. M., & Zhu, H. (2020, April 23). Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing System.*, Pg. 1 - 14. ACM. https://doi.org/10.1145/3313831.3376813

Wieringa, M. (2020). What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 1-18).

Wisconsin Supreme Court. (2016). *State v. Loomis*, 881 N.W.2d 749.

Woods, S. A., Ahmed, S., Nikolaou, I., Costa, A. C., & Anderson, N. R. (2020). Personnel selection in the digital age: A review of validity and applicant reactions, and future research challenges. *European Journal of Work and Organizational Psychology*, 29(1), 64-77.

Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J. (2018). *A Qualitative Exploration of Perceptions of Algorithmic Fairness*. https://doi.org/10.1145/3173574.3174230

Wu, Y., Zhang, L., & Wu, X. (2019, August). Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the twenty-eighth international joint conference on Artificial Intelligence*.

Xivuri, K., & Twinomurinzi, H. (2023, July 17). How AI developers can assure algorithmic fairness. Discover Artificial Intelligence, Vol 3(27), Pg. 1 - 7. Springer link. https://doi.org/10.1007/s44163-023-00074-4

Yeung, K. (2017). Hypernudge: Big Data as a Mode of Regulation by Design. *Information, Communication & Society*, 20(1), 118-136.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017, April). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th international conference on world wide web (pp. 1171-1180).

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, May). Learning fair representations. In International conference on machine learning (pp. 325-333). PMLR.

Zhang, J., & Bareinboim, E. (2018, April). Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).

Appendices

Table 8:Organizational factors influencing fairness identified from the selected articles

Theme	**Factors**	**References**
Governance	Ethical policy	Raji *et al.,* 2020; Rana *et al.,* 2023; Shneiderman 2021; Ueda *et al.,* 2023; Veale *et al.,* 2018; Xivuri *et al.,* 2023.
	Audit	Holstein *et al.,* 2019; Koefer *et al.,* 2023; Raji *et al.,* 2020; Rana *et al.,* 2023; Shneiderman 2021.
	Resource Allocation (Responsible distribution)	Adensamer *et al.,* 2021; Curto and Comim 2023; Ueda *et al.,* 2023; Veale *et al.,* 2018.
	Leadership commitment	Shneiderman 2021.
	Fairness Metrics	Ferrara *et al.,* 2023; Lee and Singh 2021; Lepri *et al.,* 2018; Madaio *et al.,* 2020; Rana *et al.,* 2023.
	Organizational Framework (Responsible Organization)	Cramer *et al.,* 2018; Holstein *et al.,* 2019; Madaio *et al.,* 2020; Madaio *et al.,* 2022; Marabelli *et al.,* 2021; Park *et al.,* 2022; Xivuri *et al.,* 2023.
	Managing stakeholders expectation	Veale *et al.,* 2018.
	Performance metrics	Madaio *et al.,* 2022. Metcalf *et al.,* 2021; Veale *et al.,* 2018.
Social Responsibility	Collaboration	Curto and Comim 2023; Cramer *et al.,* 2018; Ferrara *et al.,* 2023; Holstein *et al.,* 2019; Lepri *et al.,* 2018; Madaio *et al.,* 2020; Madaio *et al.,* 2022; Marabelli *et al.,* 2021; Madaio *et al.,* 2022. Metcalf *et al.,* 2021; Park *et al.,* 2022; Rana *et al.,* 2023; Shneiderman 2021; Starke *et al.,* 2022; Ueda *et al.,* 2023; Veale *et al.,* 2018; Xivuri *et al.,* 2023.
	Human Judgement	Aysolmaz *et al.,* 2021; Kochling and Wehner 2020; Marabelli *et al.,* 2021; Park *et al.,* 2022; Shneiderman 2021; Starke *et al.,* 2022; Wang *et al.,* 2020; Veale *et al.,* 2018;
	Sociotechnical approach	Park *et al.,* 2022.

Technical (Model Building)	Data Management	Corbett-Davies *et al.,* 2017; Curto and Comim 2023; Cramer *et al.,* 2018; Ferrara *et al.,* 2023; Holstein *et al.,* 2019; Lepri *et al.,* 2018; Marabelli *et al.,* 2021; Martin 2018; Rana *et al.,* 2023; Srinivasan and Chander 2021; Starke *et al.,* 2022; Ueda *et al.,* 2023; Veale *et al.,* 2018; Xivuri *et al.,* 2023.
	Problem solving skill	Curto and Comim 2023; Cramer *et al.,* 2018; Holstein *et al.,* 2019; Rana *et al.,* 2023; Srinivasan and Chander 2021;
	Procedural fairness	Kochling and Wehner 2020; Starke *et al.,* 2022.
	Implementation Procedure	Veale *et al.,* 2018;
	Feature selection	Corbett-Davies *et al.,* 2017; Ferrara *et al.,* 2023; Lepri *et al.,* 2018; Park *et al.,* 2022; Rana *et al.,* 2023; Srinivasan and Chander 2021; Veale *et al.,* 2018;
	Monitoring and assessment	Curto and Comim 2023;
	Tradeoff	Pessach 2022;
	New methodology approach (Substantive)	Green 2022.
	Model Testing	Xivuri *et al.,* 2023.
	Model review	Xivuri *et al.,* 2023.
Training and Development	Knowledge management	Curto and Comim 2023; Cramer *et al.,* 2018; Holstein *et al.,* 2019; Ueda *et al.,* 2023; Xivuri *et al.,* 2023.
	Communication	Aysolmaz *et al.,* 2021; Curto and Comim 2023; Holstein *et al.,* 2019; Kochling and Wehner 2020; Veale *et al.,* 2018.
	Explainability	Curto and Comim 2023; Raji *et al.,* 2020; Rana *et al.,* 2023; Shneiderman 2021; Ueda *et al.,* 2023; Veale *et al.,* 2018; Xivuri *et al.,* 2023.
	Accountability	Lepri *et al.,* 2018; Martin 2018; Madaio *et al.,* 2022. Metcalf *et al.,* 2021; Ueda *et al.,* 2023; Veale *et al.,* 2018;

	Transparency	Aysolmaz *et al.,* 2021; Curto and Comim 2023; Kochling and Wehner 2020; Lepri *et al.,* 2018; Martin 2018; Madaio *et al.,* 2022. Metcalf *et al.,* 2021; Park *et al.,* 2022; Raji *et al.,* 2020; Rana *et al.,* 2023; Ueda *et al.,* 2023; Wang *et al.,* 2020; Veale *et al.,* 2018; Xivuri *et al.,* 2023.

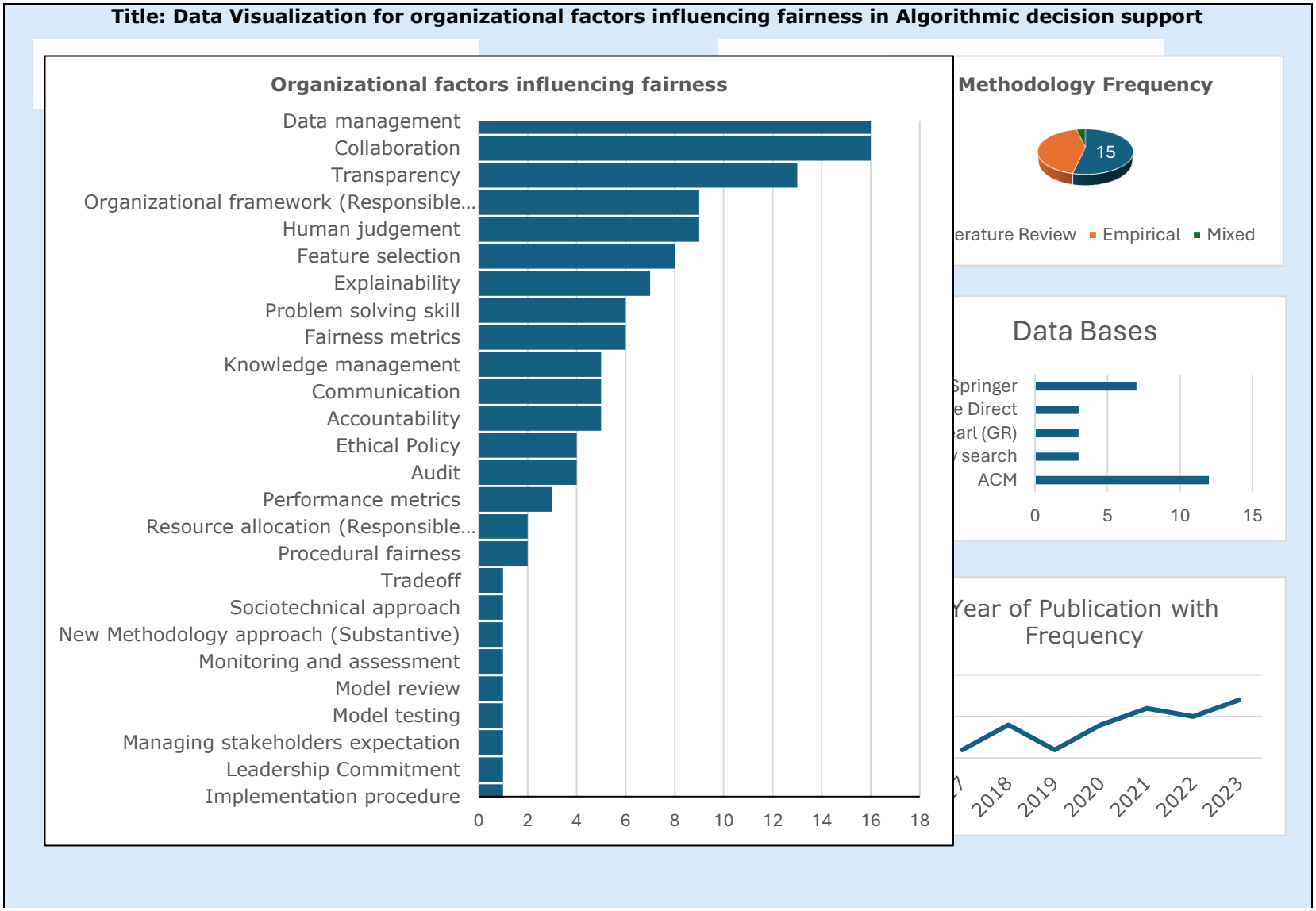

Figure 8: Data Visualization of the articles reviewed