RWTH Aachen University

Master thesis presented for the degree of Master of Science

Computational Social Systems

First Examiner: Prof. Dr. Jan-Christoph Heilinger

Second Examiner: Prof. Dr. Claudia Wagner

Towards Relational Egalitarianism in the Third Wave of Algorithmic Justice

Laila Wegner

Laila.Wegner@rwth-aaachen.de

Matriculation number: 090459

January 17, 2024

Table of Contents

1. Introduction

Machines and algorithms are precise, objective, and fact-orientated. This is the widespread assumption of machine heuristics (Sundar & Kim, 2019). Yet, ethical concerns about the use of Artificial Intelligence (AI) such as biased algorithmic decision-making (ADM) and examples of gender-biased hiring algorithms (Robert et al., 2020) or racial-biased sentencing (Angwin et al., 2016) contradict the impression of neutral and objective algorithms[1]. On the contrary, biased ADM can even reinforce stereotypes from society (Barocas & Selbst, 2016). Given that ADM is increasingly used for high-stakes and far-reaching decisions, biased ADM poses the risk of systematic disadvantage. As algorithmic biases mirror society's stereotypes and inequalities, those already marginalized are at risk of being particularly negatively affected by algorithmic discrimination (Barocas et al., 2023).

This underscores the urgency to discuss the societal impacts of ADM and AI. As a response, an interdisciplinary research area has developed, aiming to systematically address and measure the fairness of an algorithm with mathematical methods (Kasirzadeh, 2022). While the resulting fairness metrics are roughly linked to an egalitarian approach, i.e., they are based on the idea that all people are equal in some way, the contributions often lack an in-depth understanding of the moral foundations of ethical philosophy (Binns, 2018; Lee et al., 2021).

Investigating the philosophical discussion of justice, two families of approaches stand out (Arneson, 2013): On the one hand, there are distributive approaches, focusing on different currencies of equality (e.g., income, wealth, resources) and how they ought to be distributed. On the other hand, distributive theories are opposed by relational accounts, which conceptualize equality based on the quality of social relations among citizens and the treatment of citizens by social institutions. Investigating the approaches to algorithmic justice, it becomes apparent that the majority focuses exclusively on the first family of justice, i.e., they are grounded in a distributive perspective (Kasirzadeh, 2022).

However, the research on algorithmic fairness is constantly changing and can be roughly divided into three waves (Häußermann & Lütge, 2022; Huang et al., 2022): The 1st wave was

[1] The terms AI, ADM, and algorithms pose a certain challenge. While the exact definition of AI is not straightforward, the words 'artificial' and 'intelligence' are both criticized as misleading (Crawford (2021)). However, the term is widely used in the literature. The term algorithm is mostly used as a generic term for (computational) problem-solving approaches based on a defined set of rules. In the context of fairness metrics, the discussion often relies on ADM, which refers broadly speaking to classifications used to predict future behaviour or events. While this topic would deserve more attention, this would become to distinct work on its own. Therefore, I decided to use these terms interchangeably in the course of this thesis.

dominated by guidelines and principles that demanded to ensure fair development and use of ADM. To account for these guidelines, the 2nd wave of research made considerable efforts to develop mathematical solutions, identifying and mitigating unfair biases. Yet, these technical fairness metrics indicated several conceptual problems. The critique evoked the 3rd wave of AI fairness, which emphasizes that algorithms have to be seen as socio-technical systems. While the 2nd wave is rather concerned with distributive questions of the decision outcome (e.g., What is the share between men and women who received a job interview invitation?), the 3rd wave has a broader focus, including power dynamics and social structures (Kind, 2020). Consequently, the 3rd wave of algorithmic fairness extends the distributive focus toward social relations and resulting power inequalities and is therefore influenced by the thematic discourses typical of relational justice.

The presented thesis develops the notion of '*relational algorithmic justice'*, highlighting that the prevailing orientation of algorithmic fairness research towards strict egalitarian principles and questions of distributions is not suitable to account for structural injustices. When it comes to high-stakes decisions that involve value-laden considerations, i.e., decisions that are not based on facts alone but must be morally justified, a relational approach to algorithmic fairness is crucial to ensure that the impacts of automated decision-making are equitable for all. Striving for a context-based discussion, implications from relational algorithmic justice are analyzed for the moral evaluation of value-laden ADM in the case study *hiring*. For this purpose, the following research question will be investigated: Which topics, including critical and constructive approaches, are discussed within the literature on relational algorithmic justice and what are its implications for automated decision-making in hiring? Accordingly, the aim of this work is twofold: On the one hand, it synthesizes the contents of the 3rd wave of algorithmic justice and its relational implications, and on the other hand, it conducts a context-specific evaluation for the case study hiring. To this end, the presented thesis follows an interdisciplinary method based on a systematic literature review (SLR) supplemented by normative reasoning and is structured as follows:

Chapter 2 introduces the background of justice in more detail. After some terminological clarification in *Section 2.1*, the philosophical discourses on distributive and relational egalitarianism are illustrated in *Section 2.2*. Furthermore, the normative foundation of the presented thesis, based on relational justice with a special commitment towards the unconditional appreciation of human diversity, is explicated. The chapter concludes with a brief overview of the current approaches to algorithmic fairness in *Section 2.3.*

This background prepares for the analysis of the 3rd wave of algorithmic justice and its conceptual overlaps with relational justice in *Chapter 3*. Using a systematic literature review, the topics discussed in the scientific literature are identified and analyzed. Details on the method are provided in *Section 3.1* before the quantitative results are presented in *Section 3.2.* The identified topics, including constructive and critical approaches, are outlined in *Section 3.3* before the empirical part is concluded with a short methodological discussion in *Section 3.4.* The identified results will inform the philosophical discussion of implications for the use of ADM in the case study of hiring.

This is the subject of *Chapter 4.* Following a philosophical reasoning based on the commitment towards the unconditional appreciation of human diversity, it is revealed that ADM in hiring fails to recognize individual differences in *Section 4.1*. Furthermore, it will be illustrated that ADM in hiring constraints the freedom of choice in *Section 4.2.* It is concluded, that ADM in hiring fails to meet one important condition of relational equality: the appreciation of human diversity. The chapter will end with some practical recommendations in *Section 4.3* before *Chapter 5* concludes.

2. What is Justice?

As introduced before, the accumulation of examples of unfairly discriminating automated decision-making systems has led to a growing debate on justice in the context of algorithms. However, the question of justice is probably as old as humanity itself, and a common understanding and context-independent definition of justice is a difficult challenge. To disentangle the complexities at hand, the following chapter will first examine the terminological distinction between the two related concepts of justice and fairness. Afterward, this chapter illustrates the substantive questions of justice by providing a short overview of the discourse and conceptualizations of justice in the *philosophical* discourse, including the emphasis on relational accounts and the resulting commitment to human diversity. Finally, the chapter concludes with a brief introduction to the discourse of *algorithmic* justice.

2.1 Justice vs. Fairness

The plural understanding of what exactly defines justice and fairness is reflected by the different connotations of the two terms: While fairness is defined as the individual moral evaluations of rules of conduct (Goldman & Cropanzano, 2015); justice refers to a societal agreement of rules regarding a standard of rightness. These rules are mostly institutional, which means that justice has to be conceptualized within its institutional environment. This is coined by Rawls, highlighting that "justice is the first virtue of social institutions" (Rawls, 1999, p. 3). He defines justice as a question of institutional rules and practices that define rights and obligations (Rawls, 1958). The institutional dimension of justice might be contrasted by the interactional dimension of fairness. Rawls defines fairness as the rights of persons that arise in their interaction with each other on an equal basis (ibid.). In line with this, fairness may be conceptualized as decent behavior and honest attitudes toward others. Furthermore, following the distinction by Goldman and Cropanzano (2015), fairness refers to the individual evaluation of the (institutional) rules of conduct. However, the clear distinction between these terms is an ongoing controversy in the philosophical discourse.

The difference between these two terms is also easily confused within the discourse on *algorithmic justice* – or often synonymously designated as *algorithmic fairness:* One of the most influential conferences in this research area is the FAccT[2] – with the 'F' as an acronym for 'Fairness'. Also, in most of the publications and ethics guidelines, the term algorithmic fairness is more common, where fair often is used interchangeably with 'unbiased' (e.g.,

[2] ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT), https://facctconference.org/

Mehrabi et al., 2021). However, several authors reject the focus on fairness and argue for a conceptualization of algorithmic justice, which includes a more extensive engagement with social justice theories (Braun & Hummel, 2022; e.g., Bui & Noble, 2020; Green, 2018, 2022). To handle this terminological confusion, the following distinction between algorithmic justice and algorithmic fairness is used: The term algorithmic fairness describes the scientific literature that focuses on technical approaches, as this is also the term the research community uses themself. The term algorithmic justice is used in connotation of a deeper understanding of justice conceptualizations and refers to research that engages with the incorporation of philosophical fundamentals. To grip these underlying theories in more detail, it follows a closer investigation of the philosophical discourse on justice.

2.2 Justice in Philosophy

As introduced, justice is conceptualized as a societal standard of rightness and realized by institutional rules. However, this raises the substantive question of what these standards actually are. A common notion agreed upon by most modern philosophers is the idea that every person has an equal moral worth and accordingly an equal set of basic rights (Arneson, 2013). Consequently, equality is a central aspect of justice and motivated the doctrine of egalitarianism. This basic agreement leads to different implications for the derived theory of justice. In the philosophical discourse, two main accounts of egalitarianism can be overserved, which will be investigated in the following: First, *distributive* egalitarianism, which relates to questions of equal distributions of some goods, and second *relational* egalitarianism, broadly concerned with social relationships and treatment with mutual respect.

2.2.1 Distributive Egalitarianism and the Emergence of Relational Accounts

Over the decades, the philosophical discussion on egalitarianism took place within a distributive problem framework. In short, distributive justice may be summarized as theories that consider justice as the appropriate distribution of benefits and burdens among members of society. One of the fundamental questions that distinguishes distributive theories concerns the appropriate currency to analyze and measure distributive inequality. Suggested currencies include resources (Dworkin, 1981), opportunities for welfare (Arneson, 1989), or capabilities (Nussbaum, 2003). The latter builds on the distinction between *capabilities*, which are the opportunities to realize certain actions, and *functionings*, which refer to the actual fulfillment of these opportunities[3].

[3] For example, imagine two persons without a permanent home. While person A travels around the world and voluntarily decides against a permanent home, person B cannot afford to pay rent. Although both persons lack a permanent home, person A would have the capability to rent a flat but does not exercise this functioning. The fact that person B lacks the capability to rent a flat causes a moral difference in the evaluation of their situation.

Distinctions between acceptable and unacceptable distributive inequalities are often based on individual responsibility for outcomes. Following the basic assumption that being worse off without one's fault is unjust (Temkin, 1993), the societal obligation is to remove involuntary disadvantages and deliver compensation (Cohen, 1989). For example, a child born with disabilities is not to blame for the fact, that it needs more help compared to healthy children – thus, society has the moral responsibility to resolve the disadvantages caused by unfair health differences. Yet, those who caused their bad situation by bad option luck must accept their responsibility and bear the consequences of their decisions (Gosepath, 2001). This indicates two dimensions of luck: On the one hand, it exists 'bad brute luck', meaning bad luck which was not caused by those affected and demands compensation, and on the other hand it exits also the case of 'bad option luck' which is a self-inflicted bad. Situations caused by bad option luck are self-inflicted and do not justify entitlement to compensation or redistribution. This line of reasoning is also called 'luck egalitarianism', a term coined by Elisabeth E. Anderson (1999). She criticizes the focus on personal responsibility and the distinction between bad option luck and bad brute luck. In her prominent text 'What is the Point of Equality' she illustrates several cases, where individuals are excluded from compensation and welfare on the questionable justification that the bad would be their fault. For example, if the difference between bad option luck and bad brute luck is interpreted as strictly as Anderson indicates, a driver who took the risk to drive too fast and injured himself would be a case of bad option luck and accordingly, the driver would have to accept the consequences of the car accident without any entitlements to medical aid (cf. E. Anderson, 1999). This point of critique is often called the 'harshness objection' (Voigt, 2007), indicating that luck egalitarianism punishes bad choices too severely and remains indifferent to the suffering of the victims of bad option luck.

Furthermore, according to Anderson (1999), the assumption that some deserve compensation by itself relies on the assumption that some are inferior in terms of personal qualities. Remember the case of persons born with disabilities: the assumption that they need compensation directly indicates that their living conditions are worse than the conditions of healthy persons. The problem seems to be the disability itself. However, a major problem is that there are, for example, not yet good enough wheelchairs, and the public infrastructure is not sufficiently adapted to their needs. Focusing on compensation seems to indicate that those affected by disabilities are inferior and part of the problem while in fact, the societal handling of disability is the problem. Thus, as Anderson (1999) argues, compensations are based on pity, and the unfortunate are stigmatized and disrespected. Finally, she illustrates that by insisting that everyone takes personal responsibility, luck egalitarianism dictates how to make

appropriate use of individual freedom – individuals face a force to realize their optimal opportunities, always avoiding bad option luck. Additionally to Anderson's critique on luck egalitarianism, Young (1990c) emphasizes that the distributive perspective on justice tends to consider a too narrow scope: Distributive accounts fail to recognize that economic inequality is not the main source of societal injustice. Instead, it is the institutional structures in place that enable domination driven by the privileged. These critical scholars of distributive egalitarianism motivated the development of an alternative approach that focuses on relational and structural justice, concerned with social and institutional relationships. As this account will be the foundation whole thesis, the following section is devoted to relational egalitarianism in more detail.

2.2.2 Relational Egalitarianism and Structural Injustice

Relational egalitarianism is often introduced as an opponent to luck egalitarianism. Instead of referring to justice as a form of equal distribution, relational equality is concerned with the quality of social relations among individuals. According to E. Anderson (1999), the relational view demands that all people are entitled to the capabilities necessary to act as equal citizens in a democratic state. To enable equal citizenship means that every individual must have access to capabilities that influence the social order such as education or political participation. These basic capabilities function as a precondition to being an equal citizen in a democratic state. Thus, equality is rather seen as a social and political value and not as a materialistic (i.e., distributive) fact. As illustrated by the aforementioned critique on distributive accounts, relational egalitarianism identifies alternative problem fields as fundamental to egalitarian justice. Relational egalitarians consider unfair distributions as a symptom caused by social injustices. While relational egalitarians refuse distributive accounts as too limited, most of them acknowledge that relational approaches have also distributive implications. There are important distributive aspects to justice but they may not display the whole picture. For example, E. Anderson (1999) illustrates that certain distributions may exhibit instrumental values to secure relational equality. To enhance social equality some basic (distributive) needs must be met for everyone. However, relational equality demands changing the root cause (i.e., social relations and structures) and not only the symptoms (i.e., distributions) of injustices. In this spirit, Young argues that "The concepts of domination and oppression, rather than the concept of distribution, should be the starting point for a conception of social justice" (Young, 1990a, p. 16).

Oppression, domination, and unequal power hierarchies cannot be solved by distributive means only because parts of society can be still unjust even if the society has reached some kind of

distributive equality. Imagine a society where distributive inequalities have been reduced by redistribution and compensation. Someone must have decided on the strategy for redistribution and made choices about which kinds of inequality are acceptable and which are not. These choices are a form of exercising power and are probably in the hands of those privileged, maintaining unjust social practices. For example, this is observable when the global Western decides which development aid is needed in continents such as Africa. Even if the aid is well-intended and might help on a distributive level, unequal power dynamics (e.g., who has the power to decide what happens, where, and when?) and stigmatization assumptions (Africa as a needy continent dependent on the global West) are reproduced. Development aid often fails to address the causes of poverty, which are often the result of historical injustices caused by Western countries, such as exploitation in the form of colonialism.

The shifted problem framing opens the room to conceptualize justice by securing treatment with reciprocity and mutual respect. This means that no one should perceive themselves as superior or inferior to others; no one should be at the mercy of the given power structures. This applies to both the individual and face-to-face level as well as to a broader collective level in institutional contexts. Relational accounts can thus be divided into two families of relational justice (Voigt & Wester): The first one focuses on treatment with mutual respect, relating to one another as equals on the interpersonal level, for example in a partnership (see Scheffler, 2015). The second family is concerned with how institutions treat citizens. This institutional focus on relational egalitarianism was fundamentally influenced by the contributions of Iris M. Young. As will become significant in the course of this work, the institutional perspective highlights some of the most concerning issues of relational algorithmic (in)justice. Therefore, I go into more detail on Young's work in the following.

To approach relational injustices on an institutional level, Young coined the concept of structural injustices, which exist when…

> … "Social processes put large groups of persons under systematic threat of domination or deprivation of the means to develop and exercise their capacities, at the same time that these processes enable others to dominate or to have a wide range of opportunities for developing and exercising capacities available to them." (Young, 2011, p. 52)

This important quote highlights that marginalized groups are harmed while other groups gain unearned privileges and power, benefiting from structural injustices by establishing social positions with their associated advantages and disadvantages. Structural injustices become

visible in the form of racism, colonialism, sexism, and so on. All these forms of segregation disadvantage large groups in their basic needs such as finding an appropriate job, renting a flat, or getting a loan, and might expose them to exploitation and domination. Young (2011) highlights that social processes interact globally through the example of exploitative labor. By outsourcing the work to low-wage countries, dangerous, harmful, and exploitive labor conditions are reinforced – to the disadvantage of those who are dependent on the poor salary while the company and its employees benefit from it. In this way, social structures define background conditions of individual opportunities and influence the options available which either enable or constrain their actions (Young, 2006). It is these social structures that systematically benefit people in some positions by disadvantaging others and unfairly restricting their options. However, it is important to note, that individuals may contribute to structural injustices without any bad intentions. Instead, they occur due to the prioritization of the personal goals of many individuals and institutions within the societal rules and norms. For example, employees outsourcing labor to low-wage countries have to act in a highly competitive market (Young 2011) and face high pressure to make economic decisions. The employees might not intend to benefit from exploitive structures but rather strictly focus on their goal to save money. Consequently, structural injustices result from the sum of non-blameworthy actions of various actors. Hence, structural injustices cannot be reduced to individual wrongdoings and cause difficulties in attributing responsibility for the present injustices. The traditional understanding of responsibility focuses on the conditions of control and knowledge (see liability model, Young, 2006); however, both are absent given the structural nature of injustices. In summary, structural injustices are present when a sum of non-blameworthy processes limits the capabilities of large groups while others benefit from these processes and challenge a clear allocation of responsibility for the occurred harms.

Structural injustices are mainly concerned with two social constraints, namely oppression and domination (Young, 1990b). To grip this issue in more detail, Young developed the concept of "five faces of oppression" (Young, 1990b, p. 39): exploitation, marginalization, powerlessness, violence, and cultural imperialism. The latter will be particularly important for the discussion of algorithmic decision-making. Cultural imperialism describes the phenomenon in which the dominant group establishes the norm by defining other groups as the inferior deviation. The convictions of the power holders are the most widespread in society and express their experiences, values, goals, and achievements. This can be illustrated, for example, by the adoption of the language of a colonial authority or religious influences. Also, the worldwide prevailing orientation towards heteronormativity (the idea that heterosexuality is normal and

superior, and all other forms of sexuality are seen as ‘abnormal’) illustrates exemplary cultural imperialism. Cultural imperialism reduces those considered as others or abnormal to stereotypes that reflect the interpretation of the dominant group. It is also a process of rendering ‘all others’ invisible, as their experiences and perspectives are subordinated and rendered less important.

The presented theories on relational justice will serve as the foundation for both parts of the presented thesis: the systematic literature review on relational algorithmic justice (*Chapter 3)* as well as the discussion of impactions from the relational view in the case study of algorithmic decision making in hiring (*Chapter 4*). My account of relational justice is devoted to a wide interpretation of relational justice as an alternative framework to purely distributive accounts and includes both illustrated families of relational justice: The individual level, focusing on the treatment with mutual respect, as well as the institutional level, which highlights the structural nature of injustices. Furthermore, the concept of diversity and the interplay with relational justice will be essential for the argument presented in *Chapter 4.* Therefore, the following section highlights the interplay between diversity and relational justice.

2.2.3 Relational Justice and Diversity

During this chapter, I mentioned several times inequalities between different groups: those with power and those oppressed, those in the center and those in the margins, or those privileged and those harmed. The separation between those groups follows in our society in large part individual attributes such as race, gender, or social class, because, as Anderson highlights, humans misuse the different diversity dimensions to convert them into oppressive hierarchies (c.f. E. Anderson, 1999, p. 336). In this way, some diversity characteristics (e.g., gender, race) reflect patterns of status and power and are turned into oppression (e.g., sexism, racism). Focusing on the connection between relational justice and diversity, this section outlines three premises that will serve as the foundation for the practical part of *Chapter 4*. First, the unconditional appreciation of human diversity is fundamental to treating individuals *as equals* and thus required by relational justice. Premises two and three highlight that the appreciation of human diversity demands two aspects: first, the recognition of individual differences, and second, the secured freedom of choice.

1. The Unconditional Appreciation of Human Diversity is a Matter of Relational Justice

Diversity refers to the variety of human characteristics in a plural society. This includes both visible as well as invisible characteristics (Gardenswartz & Rowe, 2003), emphasizing the

individuality of each person. Considering the concept of diversity may help to understand someone's background and position within the socio-cultural structure and includes – but is not limited to – the focus on human identity, social position, or individual values (Steel et al., 2018).

Providing freedom and autonomy to express one's diversity and respecting each other's individuality is essential for supporting individual flourishing. If we ignore the diversity present, we may not truly appreciate humanity itself - humanity in its pluralistic diversity. In this spirit, Anderson stresses:

> "Egalitarianism ought to reflect a generous, humane, cosmopolitan vision of a society that recognizes individuals as equals in all their diversity. It should promote institutional arrangements that enable the diversity of people's talents, aspirations, roles, and cultures to benefit everyone and to be recognized as mutually beneficial" (E. Anderson, 1999, p. 308)

This is further illustrated by Anne Phillips (2021), highlighting that equality has to be unconditional and must not be tied to the existence of particular qualities of individuals. An unconditional understanding of equality makes no difference between the moral worth of every individual, no matter their gender, religion, or IQ – even not conditional to the human, as might be used for devaluing humans by comparisons to animal-like behavior or similar expressions (c.f. Phillips, 2021, p. 15). Unconditional equality requires the treatment *as* equals despite their differences, refusing hierarchies built on diversity dimensions such as gender, race, or "achievement-based hierarchies of education and intelligence" (Phillips, 2021, p. 4). However, the treatment *as equals* might also allow for different treatment when societal agreements are injured: For example, after the commission of a crime the person has to be treated *as* a moral equal in the judiciary, but this does not hinder the punishment of crimes (c.f. E. Anderson, 1999). To secure the treatment as equal, the offender must be supported with procedural justice during the conviction and basic human needs such as medical care and food (ibid.). Thus, the treatment as equal does not contradict the just punishment of crimes.

However, the unconditional appreciation of diversity is not yet our societal reality. The different diversity characteristics act as moderators for the social standing in our society, deciding who has the opportunity to participate in intuitional, social, or cultural environments and based on which people are discriminated against or have privileges. Given the complex dynamics of oppression in the form of racism, sexism, and so on, an intersectional analysis of diversity is needed. The term intersectionality was coined by Kimberlé Crenshaw (1989), describing how

disadvantages that occur based on individual characteristics simultaneously reinforce each other. A Black woman experiences unique forms of discrimination which cannot be analyzed by the disconnected focus on the experiences of women or the experiences of Black people (ibid.). Every individual is influenced by multiple interconnected diversity dimensions. Because these characteristics are moderators of oppressive hierarchies, intersectionality has to be analyzed within the social structures of power, too (Crenshaw, 1989). This makes the unconditional appreciation of human diversity and its intersectional dimensions fundamental for relational justice.

2. Appreciation of Diversity Demands the Recognition of Individual Differences

In the following, it will be illustrated that committing to the unconditional appreciation of human diversity demands recognizing individual differences. Following a daily understanding, recognition is associated with the act of being identified, noticed, or seen but also to be accepted. The latter is connotated to the meaning of recognition as "the act of *acknowledging* or *respecting* another being, such as when we 'recognise' someone's status, achievements or rights" (McQueen, 2023, emphasis added). This understanding is also reflected within the philosophical discourse of recognition, which can be divided into a normative and a psychological dimension (Iser, 2013).

The normative act of respecting another being includes recognizing the equal normative status and equal dignity of the other person (Iser, 2013). This illuminates the relational character of the normative force of recognition (ibid.): political movements such as 'Black Lives Matter' or pride parades such as Christopher Steet Day do not demand some kind of redistribution. Rather, they fight for the acceptance – or recognition – of their identity. Thus, the normative discourse of recognition of human diversity has substantiative overlaps with the realization of relational justice. For the normative recognition of human diversity, it is furthermore important, that different capabilities are taken into account (Robeyns & Byskov, 2011). By the acknowledgment of different capabilities, the variety of human goals and actions are recognized, as well as the impact of certain diversity dimensions on the resulting capabilities and functions. Thus, recognizing diversity includes taking different social starting points and the social environment of individuals into account, which impact their societal standing.

On the psychological dimension, it stands out that the human identity is influenced by social recognition, because it impacts self-confidence, self-respect, and self-esteem, (McQueen, 2023). Following Honneth (1995) these self-relationships are shaped by three different patterns

of recognition: the experience of love, which refers to emotional attention and positive relationships, is essential for self-confidence. The pattern of rights secures the respect for the equal worth of each individual and shapes their socialization, including the development of moral responsibility, and influences individual self-respect. Finally, Honneth (1995) stresses the importance of solidarity for the development of self-esteem. Solidarity is based on the social appreciation of different achievements (ibid.). When solidarity is lacking, individuals face a social devaluation, as their unique experiences are denied. The psychological dimension of recognition has far-reaching impacts on everyone who lacks this experience. Those who are denied recognition may not see themselves and their doings as valuable (Iser, 2013), leading to obstacles to individual flourishing and personal fulfillment. Misrecognition takes form in racism or other forms of oppression by devaluating those marginalized as inferior and refusing the recognition of their unique culture, talents, and other qualities. Thus, misrecognition results in psychological damage for those marginalized and is a form of social harm that undermines a person's positive self-relation (Iser, 2013).

As illustrated, recognition is fundamental to support positive self-relationships. A lack of recognition undermines equal chances and mutual respect. It is thus a matter of relational justice that anyone can experience recognition. Finally, this section illustrated that recognition is essential for the acknowledgment and acceptance of individual differences –including awareness of different social starting points – and is therefore fundamental for the appreciation of human diversity.

3. Appreciation of Diversity Requires to Secure Freedom of Choices

As said, diversity dimensions are often misused and turned into oppressive hierarchies. In the following it will be illustrated, how oppression based on diversity dimensions leads to constrained choices, especially for those marginalized. It is concluded that to counter oppression and appreciate diversity, it is imperative to ensure the freedom of individual choices.

The interplay between freedom of choice and oppression was notably highlighted by Ann Cudd (2006), distinguishing between direct and indirect forces of oppression. Following Cudd, oppression is to be understood as "an institutionally structured harm perpetrated on groups by other groups" (Cudd, 2006, p. 26). Social groups, then, are visible in the form of different social constraints of possible choices, determined by the group membership. These social constraints are institutionally structured and result in an unfairly limited set of options for the oppressed group. While direct forces of oppression are caused by intentional actions of the dominating

group, affecting the choices of the subordinated groups, indirect forces emerge due to rational choices by members of oppressed groups which constitute the maintenance of the oppressed state. For example, a sex worker might have chosen the job by herself and so she decides to enter oppressive structures. However, several influences such as an urgent need for money combined with no professional training might have forced her into sex work indirectly. Thus, these choices are likely due to the limited set of options and social background beliefs and desires, as the oppressed learn to behave within their social group and unconsciously follow the oppressive norms.

The observations from Cudd can be further supported by the work of Young (1990b), highlighting the relationship between choices, fixed norms, and oppression in the form of cultural imperialism. As mentioned above, cultural imperialism describes a system in which norms are defined by the dominant group, and all other groups are considered inferior. This process of norm-setting hinders the free expression of human diversity. Remember the example of heteronormativity which considers all sexualities beyond heterosexuality as deviation from the norm and implies, that everyone should adapt towards the heterosexual norm. This may lead to the fact that individuals suppress their wishes or only live them out in secret. When choices are limited within systems of oppression, humans have an unfairly limited set of options available which shapes their own identity. This might also lead to the suppression of their individuality by adopting the norms established by the dominant group. Ensuring freedom of choice is therefore essential to allow individuals to develop their own identity and thus appreciate human diversity.

All in all, this section illustrated the appreciation of diversity as a fundamental issue for the conceptualization of relational justice. This appreciation demands on the one hand the recognition of individual differences while securing freedom of choice in all their variety. Summarizing the importance of appreciating diversity, Iris M. Young stressed that…

> "Social justice […] requires not the melting away of differences, but institutions that promote the reproduction of and respect for group differences without oppression" (Young 1990c, S. 47).

In this work, I will apply the previously illustrated philosophical groundings of relational justice to the context of ADM in hiring. However, this requires a basic understanding of the status quo on algorithmic justice, which I will briefly sketch in the following before I turn to the investigation of relational algorithmic justice in *Chapter 3*.

2.3 Algorithmic Justice

The research field of algorithmic justice developed with the aim to counteract algorithms that systematically disadvantage marginalized groups. As mentioned previously, the resulting research can be roughly divided into three waves (Häußermann & Lütge, 2022; Huang et al., 2022): Whereas the **first** wave focuses mainly on guidelines and principles aiming to ensure a fair development and use of ADM, the **second wave** deals with mathematical approaches to identify and mitigate unfair biases. Finally, the **third wave** focuses on the conception of algorithms as a socio-technical system, including power dynamics and social structures (Huang et al. 2022, quoting Kind, 2020). In the following, I provide a brief overview of each wave.

The **1st wave** is characterized by discussing appropriate guidelines and principles to ensure a fair development and use of algorithms or AI. Among them, the ethics guideline for trustworthy AI developed by the 'EU High-Level Expert Group' (HLEG, 2019) is one of the most popular but controversially discussed: For example, Metringer (2019) raises the criticism that these guidelines are susceptible to 'ethics washing' and highly influenced by industry stakeholders – to name just two points of critic. However, in addition to the European approach, over 80 AI ethics principles were already published in mid-2022, and the trend is rising (Munn, 2022). Jobin et al. (2019) identify five recurring key principles within the AI ethics guidelines: transparency, justice and fairness, non-maleficence, responsibility, and privacy. Yet, it is difficult to synthesize these principles in more detail, as different guidelines introduce substantive differences in the interpretation of these principles, their underlying rationale, and the questions of who should implement these principles how, and in which domain (ibid.). In turn, these guidelines have been criticized as meaningless because they lack clear specifications on how the principles should be realized in practice (Munn, 2022). This leads to the observation that many guidelines lack impact, caused of their generality and superficiality (Hagendorff, 2020).

Aiming to concretize, formalize, and implement the principle of 'justice and fairness', the **2nd wave** is characterized by technical approaches to identify and mitigate biases. In this context, a 'fair' algorithm is roughly interpreted as one that is not skewed or biased in favor of one group of people (Mehrabi et al., 2021). To narrow down which types of biases represent characteristics that are morally acceptable to discriminate on and those that are unacceptable, the approaches

are based on the concept of protected attributes[4] such as gender, race, and age which represent characteristics that should not impact the decision outcome (Caton & Haas, 2020). However, simply deleting these attributes from the data set does not work, because of many statistical correlations that indicate a relation between a protected and a non-protected attribute (Dwork et al., 2012). For example, the non-protected attribute ZIP code often correlates to the protected attribute race, so a decision that distinguishes based on ZIP codes would also distinguish based on race. This problem is often called 'proxy discrimination' and is the reason why approaches such as 'fairness through unawareness' do not work in the algorithmic context (Dwork et al., 2012). In turn, scientists developed over 20 definitions to computationally measure fairness, with none of them being the 'best' one (Narayanan, 2018). The number of different definitions introduced a trade-off in the choice of the used definition, as some of them are mathematically incompatible and address a distinct normative question that goes beyond the technical perspective. However, algorithmic fairness research tends to focus mostly on mathematical solutions and is commonly concerned with distributive questions regarding the prediction outcome (Kasirzadeh, 2022). For a comprehensive overview of the details behind these mathematical approaches to algorithmic fairness, see Caton and Haas (2020), Corbett-Davies and Goel (2018), or Pessach and Shmueli (2022), and for a detailed view of the assumptions, mathematical methods, and the and interplay with society see Barocas et al. (2023).

However, there are some fundamental characteristics that the fairness metrics have in common. The majority of fairness metrics follow a group-based approach[5]. Group-based approaches are concerned with differences within a protected group (e.g., gender), indicated by statistical imbalances across different group memberships (e.g., men and women) (Pessach & Shmueli, 2022). These statistical imbalances are mostly formalized by true positive, true negative, false positive, or false negative outcomes. A true positive refers to correctly positive classified cases, i.e., for example in hiring an applicant who is qualified and invited to the job interview. If the applicant would not be invited even though he or she would have been qualified for the job, this is an example of a false negative. Analogous, a true negative prediction represents a person who is not qualified for the job and is not invited, while a false positive represents a person who is not qualified but gets an unjustified positive prediction and is still invited for the job interview.

[4] A list of protected attributes indicates the assumption that a straightforward distinction between acceptable and unacceptable inequalities is possible, justifying the distinction between protected and unprotected attributes (see Lee et al. (2021)). Whether this is the case, is subject to discussion as we will see in *3. Systematic* Literature Review: Towards Relational Algorithmic Justice

[5] Unlike group-based fairness metrics, individual fairness looks at the outcome for each decision subject (Caton and Haas (2020)). This includes the approach of counterfactual fairness, which tests whether the outcome would change if the protected attribute had been different (ibid.)

This measurement of performance is commonly summarized in a confusion matrix, with the two dimensions of actual and predicted outcome (see *Table 1*). The confusion matrix represents the common foundation of group-based fairness metrics, which differ in the emphasis on the respective cases: For example, the metric 'equal opportunity' is based on the share of true positives between two groups, while the calculation of the metric 'equalized odds' is based on false positives and so on (Caton & Haas, 2020).

Table 1 Confusion Matrix - the common basis of group-based fairness metrics

	Predicted Negative	**Predicted Positive**
Actual Negative	True Negative	False Positive
Actual Positive	False Negative	True Positive

The question of the normative rationale behind each of these fairness metrics is most notably highlighted by Binns (2018). He points out that the existing approaches of fairness in machine learning are roughly based on philosophical ideas, but the researchers do not always explicitly provide a moral foundation. Binns (2018) cautions that algorithmic fairness research might fail to consider the justification behind the use of protected attributes, ending up in narrow approaches that do not sufficiently consider philosophical implications. A further tension on algorithmic fairness is described as the risk of a "framing trap" (Selbst et al., 2019, p. 60) which describes the misconception that as soon as a certain fairness criterion is met, the technology is considered legitimate while failing to consider the whole system over which an algorithm will be enforced.

Furthermore, algorithmic fairness relies on the underlying assumption that computational fairness may be evaluated *separately* without considering the social network of the decision subject (Mitchell et al., 2021). For example, in the algorithmic fairness literature, the fairness of a loan decision is only evaluated based on the individual decision subject (ibid.). The algorithmic fairness evaluation may neither take into account whether or not the loan has an impact on the subject's family, nor whether or not the decision subject has a social network that could help in case of a loan denial. Influences from the social network on the urgency of obtaining a loan cannot be considered in a separate evaluation of each decision subject. Another unconscious assumption is the idea that the fairness of a given prediction can be evaluated *symmetrically* among group members (ibid.). This would mean, that – considering the protected attribute of (binary) gender – the fairness of a loan denial and the resulting harm would be the same for each woman or respectively men; without considering why the loan was needed and that a loan denial is likely to have different impacts for each individual. A further weakness of statistical fairness metrics is highlighted by Barocas et al. (2023) pointing out that statistical

fairness metrics are not able to identify structural discrimination which is reflected in the data: For example, education level may be a relevant feature to consider for a job position but correlates with the socioeconomic background of the applicants. By defining structural discrimination as arising from "ways in which society is organized, both through relatively hard constraints such as discriminatory laws and through softer ones such as norms and customs" (Barocas et al., 2023, p. 209), they already point closely to the theories of structural injustices as introduced in the previous section, even if they do not directly engage with the philosophical literature.

The previous paragraph illustrates exemplarily, that the statistical fairness metrics received more and more critique because of several conceptual problems. The critique on the purely technical approaches evoked the **3rd wave** of AI fairness, which emphasizes that algorithms must be seen as socio-technical systems. Häußermann and Lütge (2022) highlight that this third wave is currently developing through critical contributions to the previous waves and does not yet have a clear upshot. They furthermore highlight that this wave should strongly engage with normative goals with a thorough theoretical basis to develop appropriate practical implementations. In general, scholars of the 3rd wave of AI ethics demand to consider the sociotechnical nature of algorithms and to be sensitive to social contexts where an algorithm is applied (Huang et al., 2022; Kind, 2020). The 3rd wave of AI ethics was named for the first time by Carley Kind, director of the Ada Lovelace Institute, in a post-blog-post which is increasingly taken up by scientific literature (e.g., Borg, 2021; Braun & Hummel, 2022; Burr & Leslie, 2023; Huang et al., 2022). She describes the 3rd wave as follows:

> "Third-wave ethical AI is less conceptual than first-wave ethical AI and is interested in understanding applications and use cases. It is much more concerned with power, alive to vested interests, and preoccupied with structural issues, including the importance of decolonising AI." (Kind, 2020)

This quote indicates that the 3rd AI ethics wave is characterized by topics within the discourse of relational egalitarianism. Similar to relational egalitarianism, the third wave demands to overcome the exclusively distributive focus of algorithmic justice. In this spirit, Branford (2023) stated at the CEPE[6] that AI Ethics currently face a "relational turn" which should be further encouraged.

[6] International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023

Within the next chapters, I investigate what the relational perspective adds to the – *so far merely distributive* – discourse of algorithmic justice. However, the following discussion on algorithmic justice is not to say that we shouldn't care for the resulting distributions of an ADM. An unbiased algorithm might be desirable in certain contexts, but it also does not serve as a complete solution to enable algorithmic justice. While acknowledging that fair distributions of algorithmic predictions are important, I stress that it is not the whole picture. What this whole picture might look like and how topics such as the impact of algorithms on structural injustices, power dynamics, or other broader societal issues are addressed in the 3rd wave of algorithmic justice will be investigated in the systematic literature review in the following chapter. There are to date and to the best of my knowledge no systematic reviews corresponding to the 3rd wave. To conceptualize these topics for a systematic literature review, the observation of the substantive overlap between the topics discussed within the 3rd wave of algorithmic justice and relational egalitarianism will be used to identify relevant search keywords.

3. Systematic Literature Review: Towards Relational Algorithmic Justice

To address the aforementioned research gap and to provide first insights into how the 3rd, relational wave of algorithmic justice can be characterized a systematic literature review (SLR) was conducted. It aims to answer the first part of the overall research question of the presented thesis: Which topics, including critical and constructive approaches, are discussed within the literature on relational algorithmic justice? The findings will provide the basis for discussing the implications of relational algorithmic justice using the case study of ADM in hiring in *Chapter 4*.

In the following, details on the research *method* are provided, before a structural analysis of the literature is presented in the section *quantitative results*. Afterwards, the qualitative results are presented in the section *findings*, before the presented SLR concludes with a short *discussion*.

3.1 Method

To provide valuable insights, a systematic literature review must be characterized by a clearly defined method that is replicable to a certain degree. By collecting and synthesizing previous research in an unbiased way, it is possible to gain more general knowledge about the research object under consideration (Kitchenham, 2004; Snyder, 2019). Therefore, SLRs must adhere to strict guidelines for search strategy and article selection (Snyder, 2019). To meet this requirement, the following SLR is methodologically based on the well-established guidelines by Kitchenham (2004). In her work, Kitchenham recommends the following three phases to conduct an SLR: In the first phase, the planning phase, research questions are formulated, review objectives are identified and a pre-defined review protocol is developed that outlines the inclusion and exclusion criteria. In the second phase, the SLR is conducted and a quantitative description of the identified literature is presented. The SLR is finalized within the third phase: providing the qualitative findings.

After the research question was formulated, the following concise review protocol ensures a transparent and replicable process. To define an adequate search strategy, an extensive keyword search was conducted, leading to a search query that consisted of three parts: first, a technical keyword such as algorithm, artificial intelligence, machine learning, or algorithmic decision combined with *second*, the broad topic of (in-)justice and (un-)fairness and *third*, a specific relational justice keyword which includes the following terms: structural injustice, structural

justice, feminism, feminist, critical theories, social hierarchies, social status, social class, power dynamics, human power, power hierarchies, participatory, democratization, relational ethics, relational egalitarianism, relational equality, domination, and oppression. Each part is combined with the Boolean operator AND, so that at least one keyword per part must be included in the title, keywords, or abstract of the relevant literature. The final search queries adapted to each database are depicted in *Table 2*.

However, to obtain an unbiased and reproducible selection of the keywords, a transparent rationale for the choices made is needed (Snyder, 2019) which is summarized in the following. To identify the search keywords for relational justice comprehensively, Nath's (2020) article on relational egalitarianism in the journal Philosophy Compass, a well-established journal containing survey articles on current topics in philosophy, was used for a first orientation. Nath illustrates the diverse nature of the discourse on relational egalitarianism and its key issues. Especially objections against hierarchies stand out: Some relational egalitarians reject unequal social status hierarchies and others focus on unequal power dynamics. To approach papers concerned with unequal social status hierarchies I included the keywords "social hierarchies", "social status" and "social class". Because the term 'power' often relates in technical papers to electricity or performance, I specified the issue of power by the keywords "human power", "power dynamics" and "power hierarchies". Unequal hierarchies may undermine relational equality because they enable domination (Nath, 2020) and oppression which are at odds with the idea of people respecting each other as equals. Therefore, I added "oppression" and "domination" to my search query. The disapproval of oppression also refers to oppressive states, thus relational egalitarianism rejects undemocratic governments (ibid.). From this, a high value of democratic processes can be derived and I decided to include democratization and the related concept of participation into my query. As mentioned in *Section 2.2*, the institutional understanding of relational justice includes particular attention to structural injustice. Therefore, I included the terms "structural justice" and "structural injustice". Nath (2020) furthermore highlights that relational equality is also expressed by feminist and critical theories so I included the terms "feminism", "feminist" and "critical theories". Finally, I included broader terms that indicate the reference to relational egalitarianism (i.e., "relational ethics", "relational egalitarianism" and "relational equality") in my search query.

To ensure that the SLR captures the variety of the interdisciplinary research field, I deployed the described search query on the multidisciplinary search databases (Web of Science and Scopus) as well as relevant discipline-specific databases (ACM Digital Library and PhilPapers).

Table 2 Search query per database and excluded research areas

Database	Search Query	Excluded Research Areas
Web of Science	(TS=(Algorithm*) OR TS=("Artificial Intelligence") OR TS=("Machine Learning") OR TS=("Automated Decision")) AND (TS=(*justice) OR TS=(*fairness)) AND (TS =("Structural Injustice") OR TS=("Structural Justice") OR TS=(Feminis*) OR TS=("Critical Theories") OR TS=("Power Dynamics") OR TS=("Human Power") OR TS= ("Power Hierarchies") OR TS=(Participatory) OR TS=(Democratization) OR TS=("relational ethics") OR TS=("Relational Egalitarianism") OR TS=("Relational Equality") OR TS=(oppression) OR TS=(domination) OR TS=("social hierarchies") OR TS=("social status") OR TS=("social class"))	Engineering; Physics; Chemistry; General Internal Medicine; Instruments Instrumentation; Biochemistry Molecular Biology; Construction Building Technology; Materials Science; Meteorology Atmospheric Sciences; Neurosciences Neurology; Pharmacology Pharmacy; Rheumatology
Scopus	TITLE-ABS-KEY((Algorithm* OR “Artificial Intelligence” OR “Machine Learning” OR “Automated Decision”) AND (*justice OR *fairness) AND (“Structural Injustice” OR “Structural Justice” OR feminis* OR “critical theories” OR “Power Dynamics” OR “Human Power” OR “power hierarchies” OR Participatory OR Democratization OR “relational ethics” OR “relational egalitarianism” OR “relational equality” OR oppression OR “domination” OR “social hierarchies” OR “social status” OR “social class”))	Engineering; Energy; Physics and Astronomy; Biochemistry, Genetics and Molecular Biology; Neuroscience; Pharmacology, Toxicology and Pharmaceutics; Materials Science; Chemistry; Chemical Engineering; Agricultural and Biological Sciences; Economics, Econometrical and Finance
ACM Digital Library	(Abstract:(Algorithm* OR "Artificial Intelligence" OR "Machine Learning" OR "Automated Decision") OR Title:(Algorithm* OR "Artificial Intelligence" OR "Machine Learning" OR "Automated Decision") OR Keyword:(Algorithm* OR "Artificial Intelligence" OR "Machine Learning" OR "Automated Decision")) AND (Abstract:("justice" OR "fairness" OR "injustice" OR "unfairness") OR Title:("justice" OR "fairness" OR "injustice" OR "unfairness") OR Keyword:("justice" OR "fairness" OR "injustice" OR "unfairness")) AND (Abstract:("Structural Injustice" OR "Structural Justice" OR feminis* OR "critical theories" OR "Power Dynamics" OR "Human Power" OR "power hierarchies" OR "Participatory" OR "Democratization" OR "relational ethics" OR "relational egalitarianism" OR "relational equality" OR "oppression" OR "domination" OR "social hierarchies" OR "social status" OR "social class") OR Title:("Structural Injustice" OR "Structural Justice" OR feminis* OR "critical theories" OR "Power Dynamics" OR "Human Power" OR "power hierarchies" OR "Participatory" OR "Democratization" OR "relational ethics" OR "relational egalitarianism" OR "relational equality" OR "oppression" OR "domination" OR "social hierarchies" OR "social status" OR "social class") OR Keywords:("Structural Injustice" OR "Structural Justice" OR feminis* OR "critical theories" OR "Power Dynamics" OR "Human Power" OR "power hierarchies" OR "Participatory" OR "Democratization" OR "relational ethics" OR "relational egalitarianism" OR "relational equality" OR "oppression" OR "domination" OR "social hierarchies" OR "social status" OR "social class"))	The database focuses on Computing Literature; no further filters exist
Phil Papers	(Algorithm* | "Artificial Intelligence" | "Machine Learning"| "Automated Decision") & (*justice | *fairness) & … … ("structural injustice" | "structural justice"); … (feminis* |oppression); … (domination | "critical theories"); … ("human power" |"power dynamics"); … ("power hierarchies "); … (Participatory | Democratization); … ("relational ethics"); … ("relational egalitarianism"); … ("relational equality"); … ("social hierarchies"); … ("social status" | "social class")	The database focuses on philosophical literature; no further filters exist

While PhilPapers is included because of its domain-specific nature, the choice of the three other databases follow the recommendation to access databases with high precision, recall, and reproducibility (Gusenbauer & Haddaway, 2020). The multidisciplinary databases are very large, as Scopus includes more than 90 million peer-reviewed entries (Elsevier, 2023), and Web of Science covers over 89 million records (Clarivate, 2023). ACM Digital Library includes over 3.5 million entries (ACM, 2023). PhilPapers with roughly 2.8 million entries (The PhilPapers Foundation, 2023) is the smallest of the four databases. ACM Digital Library, Scopus, and Web of Science allow for query strings with combined Boolean operators. While it is possible to use Boolean operators on PhilPapers, too, the search query was too long and had to be divided into several shorter search queries, as illustrated in *Table 2*.

Pre-defined inclusion and exclusion criteria supplemented the search strategy. A limitation of the publication period was not included to analyze the development of the publications over time. Studies that fulfilled at least one of the following criteria were, if possible, directly excluded by filters in the search engines:

1. The study is not written in English.
2. The study is not an academic publication – i.e., only articles in peer-reviewed journals and conference proceedings were included.
3. The study is assigned to a clearly unrelated research area such as astronomy or physics. All excluded research areas are listed in *Table 2*.

After the search queries had been run, the documents were exported and duplicates were removed, followed by the manual study selection. To this end, I performed several screening rounds based on title, abstract, or full text. In the first round, titles that indicate a different topic have been removed. Afterwards, I pre-formulated three questions which were used as a guide throughout the remaining screening rounds:

1. Does the study engage with a deeper understanding of structural (in)justice or is referring to approaches from relational egalitarianism?
2. Is the paper not limited to a specific and narrow research area but focuses on broader societal issues?
3. Does the paper directly discuss algorithms/ machine learning/ AI and is not referring to technological phenomena in general?

Documents that did not answer these questions positively were excluded. All documents that remained after these screening rounds were subjected to an in-depth text analysis.

3.2 Quantitative Results

The keyword-based literature search was conducted on the 25th of September, 2023. In the following, I will outline the descriptive results of the review process. Furthermore, each step of paper selection is depicted in *Figure 1* as recommended in several SLR guidelines (Joshua D. Harris et al., 2013; e.g., Kitchenham, 2004; Page et al., 2021).

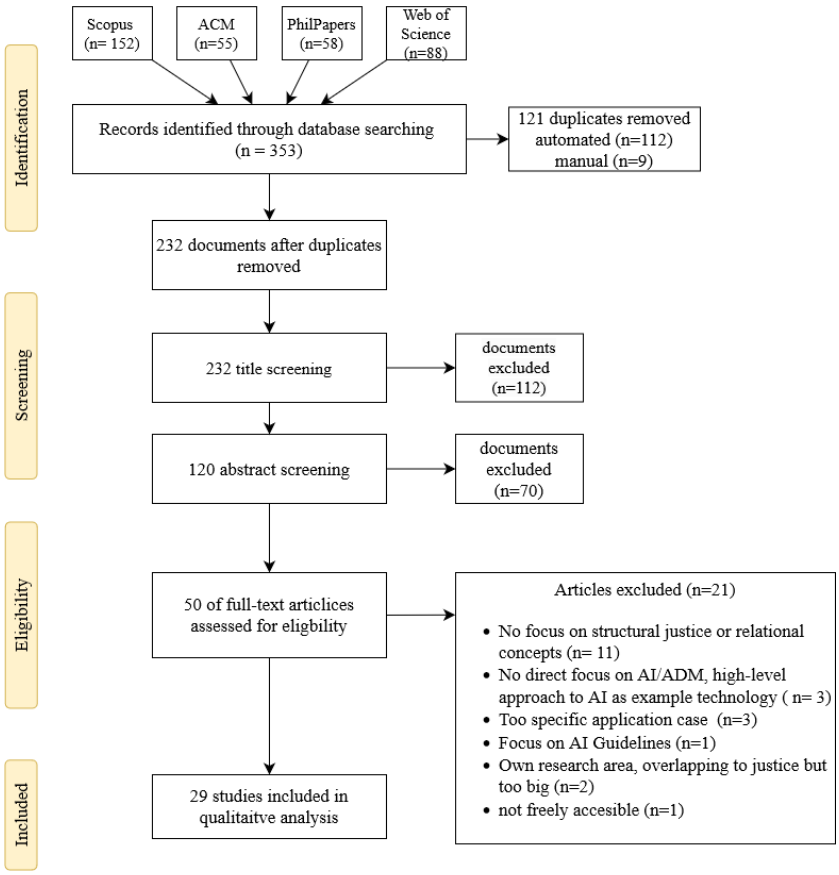

Figure 1 Flowchart illustrating the process of the SLR

Requesting the described search queries from four databases resulted in **353 documents** that were exported to a CSV or XLS document. In the next step, all titles were preprocessed to lowercase and punctuation removed, before the pandas function 'drop_duplicates' was used on the column titles to delete duplicates. The full code is included in Appendix A: Preprocessing. In total, 112 duplicates were removed automatically and nine duplicates manually in cases when two different titles referred to the same paper (e.g., long and short form), so **232 documents** remained for the screening procedure.

Afterward, titles, abstracts, and full texts of all remaining documents were reviewed in several rounds to determine the articles' relevance to my research scope. In the first screening phase, only the title was used to exclude papers that were clearly out of the scope (e.g., participatory budgeting; power dynamics in the context of the energy sector and not related to human power)

and those allocated to a specific application case out of my research scope (e.g., bias in natural language processing). Furthermore, two secondary sources (reviews) were excluded. By title screening 112 papers were discarded, leaving **120 documents**. Next, I screened all remaining abstracts and excluded documents which are not peer-viewed (e.g., books) and those out of my research scope. For example, some authors mention the use of an algorithm or artificial intelligence but do not discuss the functioning of the technology itself. Other articles mentioned justice or fairness as a broader motivation for their research, but it was not their core topic. Abstract screening resulted in the rejection of 70 articles so **50 documents** were left for full-text screening. In this last screening phase, further 16 documents were excluded because, for example, they did not engage with a relational or structural perspective on algorithmic justice or AI ethics; or because they did not discuss AI/ ADM directly but worked with a high-level understanding of AI as technology in the broader sense. Furthermore, five more papers were excluded after detailed reading. The full list of reasons to exclude documents in the eligibility assessment is depicted in *Figure 1*. In summary, this procedure based on predefined inclusion and exclusion criteria resulted in **29 documents** that remained for the final selection. *Table 3* shows the complete list of the 29 selected papers. Although the search was not limited to a time frame, there is only one paper from 2017 and one from 2019 – the rest were published in 2020 or even more recently. As illustrated in *Figure 2*, the chronological development of the identified literature underscores the novelty and increasing importance of research on structural algorithmic justice in the last couple of years. The observation that only one paper is from 2023 may be because the research was conducted in September and thus not all papers of the recent year are online yet.

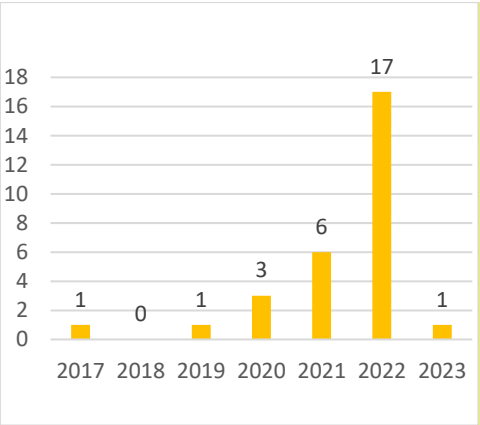

Figure 2: New Paper contributions per Year

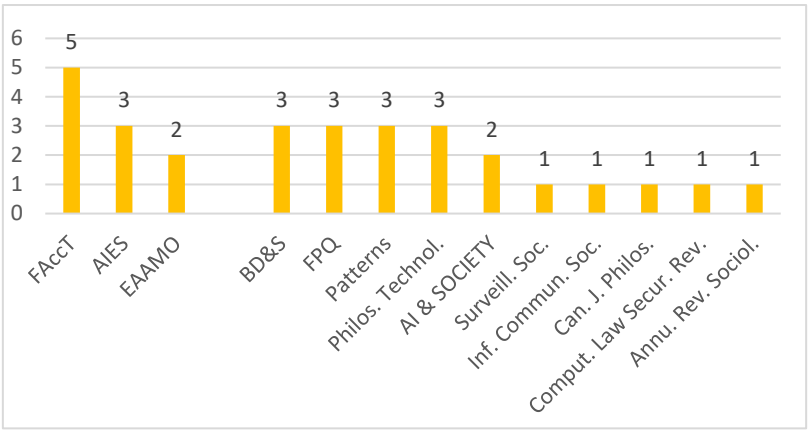

Figure 3: Distribution of Publications. Left: conferences, right: journals

10 papers are published as conference proceedings, where the ACM Conference on Fairness, Accountability, and Transparency (FAccT) is known for a high conference rank and is very influential in the field of algorithmic justice. But also, the other two conferences – AI, Ethics, and Society (AIES) and Algorithms, Mechanisms, and Optimization (EAAMO) – are well-known. *Figure 3* illustrates the distribution of conferences or journals where the selected papers are published. 19 papers are published in journals, of which well-known journals such as Big Data and Society (BD&S), Feminist Philosophy Quarterly (FPQ), Patterns as well as Philosophy & Technology are represented with three papers each.

In terms of the national context, a relatively high homogeneity is observable. With 44,8 %, the majority of the papers are written by authors working in the USA. The United Kingdom represents the second largest portion of studies with 17,2 %, followed by Germany, Ireland, and Netherlands with 10,3% each. With one paper from Canada and one paper from Hong Kong, these countries contribute 3,4% of the total collection. All in all, it should be noted that this collection represents almost exclusively Western democracies.

Table 3 Overview of papers included

Authors	Year	Title	Published in	Country
Aizenberg E.; van den Hoven J.	2020	Designing for human rights in AI	Big Data & Society	Netherlands
Andrus M.; Villeneuve S.	2022	Demographic-Reliant Algorithmic Fairness: Characterizing the Risks of Demographic Data Collection in the Pursuit of Fairness	FAccT	USA
Birhane A.	2021	Algorithmic injustice: a relational ethics approach	Patterns	Ireland
Birhane A.; Isaac W.; Prabhakaran V.; Diaz M.; Elish M.C.; Gabriel I.; Mohamed S.	2022	Power to the People? Opportunities and Challenges for Participatory AI	EAAMO	Ireland
Braun M.; Hummel P.	2022	Data justice and data solidarity	Patterns	Germany
Burrell J.; Fourcade M.	2021	The Society of Algorithms	Annual Review of Sociology	USA

Cinnamon J.	2017	Social injustice in surveillance capitalism	Surveillance & Society	UK
Gangadharan, SP; Niklas, J	2019	Decentering technology in discourse on discrimination	Information, Communicatio n & Society	UK
Green B.	2022	Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness	Philosophy & Technology	USA
Gwagwa A.; Kazim E.; Hilliard A.	2022	The role of the African value of Ubuntu in global AI inclusion discourse: A normative ethics perspective	Patterns	Netherlands
Hampton LM	2021	Black Feminist Musings on Algorithmic Oppression	FAccT)	USA
Heilinger J.-C.	2022	The Ethics of AI Ethics. A Constructive Critique	Philosophy & Technology	Germany
Himmelreich J.	2023	Against "Democratizing AI"	AI & SOCIETY	USA
Huang, Linua Ta-Lun; Chen, Hsiang-Yun; Lin, Ying-Tung; Huang, Tsung-Ren; Hun, Tzu-Wie	2022	Ameliorating Algorithmic Bias, or Why Explainable AI Needs Feminist Philosophy.	Feminist Philosophy Quarterly	Hong Kong
Kasirzadeh A.	2022	Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy	AIES	UK
Kassam, Alysha; Marino, Patricia	2022	Algorithmic Racial Discrimination: A Social Impact Approach	Feminist Philosophy Quarterly	USA
Kong Y.	2022	Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis	FAccT	USA
Krupiy T.T.	2020	A vulnerability analysis: Theorising the impact of artificial intelligence decision-making processes on individuals, society and human diversity from a social justice perspective	Computer Law & Security Review	Netherlands
Leavy S.; Siapera E.; O'Sullivan B.	2021	Ethical Data Curation for AI: An Approach based on Feminist Epistemology and Critical Theories of Race	AIES	Ireland
Lin, Ting-An Lin; Chen, Po-Hsuan Cameron	2022	Artificial Intelligence in a Structurally Unjust Society.	Feminist Philosophy Quarterly	USA
Lu C.; Kay J.; McKee K.	2022	Subverting machines, fluctuating identities: Re-learning human categorization	FAccT	UK
Rafanelli L.M.	2022	Justice, injustice, and artificial intelligence: Lessons from political theory and philosophy	Big Data & Society	USA
Sloane M.; Moss E.; Awomolo O.; Forlano L.	2022	Participation Is not a Design Fix for Machine Learning	EAAMO	USA
Tacheva Z.	2022	Tracking a critical look at the critical turn in data science: From "data feminism" to transnational feminist data science	Big Data & Society	USA
Tomasev N.; McKee K.R.; Kay J.; Mohamed S.	2021	Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities	AIES	UK
Wong P.-H.	2020	Democratizing Algorithmic Fairness	Philosophy & Technology	Germany
Zajko, Mike	2021	Conservative AI and social inequality: conceptualizing alternatives to bias through social theory.	AI & SOCIETY	Canada
Zhang M.	2022	Affirmative Algorithms: Relational Equality as Algorithmic Fairness	FAccT	USA

3.3 Findings

The following presentation of the qualitative findings is divided into two sections: Given the novelty of analyzing algorithmic justice from a relational perspective, the following section will first outline its *underlying motivation* and broader criticism of the algorithmic justice research. It follows the presentation of the distinct topics in terms of *critical and constructive approaches* which are derived from a relational perspective on algorithmic justice before the findings are discussed in the next section (see *3.4 Discussion*).

3.3.1 Underlying Motivation

First of all, the SLR has shown, that the discussion of structural injustices and relational accounts has indeed arrived in the algorithmic justice literature. In general, there seems to be a gap between algorithmic 'fairness' and the concepts of fairness and justice in the philosophical sense, as several authors introduced their terms for delimitation between fairness in the philosophical sense and algorithmic fairness, such as "formal algorithmic fairness" (Green, 2022; Zimmermann & Lee-Stronach, 2021), the "dominant interpretation of [intersectional] fairness" (Kong, 2022) or the "dominant approach to AI fairness." (Lin & Chen, 2022) as well as speaking from "fairness-as-parity conceptualizations" (Kassam & Marino, 2022). Furthermore, Hampton (2021) and Zajko (2021) criticize that the technical language – i.e., speaking about 'algorithmic bias' and 'unequal ground truth' – fails to capture the dimension of social justice and structures of oppression. Therefore, it is needed to replace the wording of 'biases' with a more nuanced grammar and vocabulary from theories of social justice (Cinnamon, 2017; Rafanelli, 2022; Zajko, 2021).

The critical observation that algorithmic fairness focuses on narrow and technical approaches and fails to account for structural injustices is a common starting point of several authors: Wong (2020) denotes the current approaches to algorithmic fairness as unsatisfactory because they are "primarily conceptualized as a technical challenge" (p. 226) which may not account for the political dimension of fairness. In a similar vein, Kasirzadeh (2022) illustrates that the methodology of the current algorithmic fairness research is to propose mathematical metrics. Furthermore, several other authors claim that the technical solutions are too simplistic, neglecting a deeper problem understanding of the societal context (Aizenberg & van den Hoven, 2020; Birhane, 2021; Gangadharan & Niklas, 2019; Green, 2022; Heilinger, 2022; Zhang, 2022). Following Zimmermann and Lee-Stronach (2021) the like cases maxim, which is a necessary condition for procedural justice and demands that individuals with similar moral characteristics must receive equal treatment, is violated by algorithms because they fail to

consider background conditions of structural injustice (ibid.). However, these conditions are moderators of the algorithmic prediction (e.g., through proxy discrimination), impacting the decision outcome (ibid.). A similar observation is made by Rafanelli (2022), who claims that algorithmic bias contributes to structural injustices by facilitating unintentional and institutional discrimination. Birhane (2021) highlights that injustices "are treated as side effects that can be treated with technical solutions" (p. 2), which fail to account for asymmetrical power hierarchies. Similarly, Kassam and Marino (2022) stress that algorithmic justice has to emphasize the asymmetrical and hierarchical nature of racism. Furthermore, Birhane (2021), Heilinger (2022), and Green (2022) explicitly demand to acknowledge relational considerations. Derived from the unequal power dynamics and unjust background conditions, the need to directly address algorithmic oppression is highlighted by Hampton (2021), Leavy et al. (2021), and Braun and Hummel (2022).

These observations led to the conclusion that the research on algorithmic fairness is restricted to distributive ideals, as statistical fairness metrics aim for equal distributions (Kasirzadeh, 2022; Kong, 2022; Lin & Chen, 2022). Kong (2022) criticizes this focus because in fact "it is misleading to say that an algorithm's fairness depends only on how it distributes the number" (p. 491). There is a consensus that the current approaches of algorithmic fairness only address the outcomes, which represent the consequences of structural injustices, while it would be necessary to address the root causes of social injustices (Heilinger, 2022; Kasirzadeh, 2022; Kong, 2022; Krupiy, 2020). As Heilinger (2022) brings to the point, "technological solutions will even in the best case provide nothing but cosmetic improvements at the symptom level, leaving the origins, structures and underlying social dynamics and structural injustices untouched" (p. 11).

This leads to the problem that many technical approaches are misguided in their roots because they overlook the fact that social bias cannot be 'fixed': This misled assumption of 'fixing' unfairness arises from the focus on debasing algorithms (Birhane, 2021; Lin & Chen, 2022), reducing systems of oppression to the discussion of algorithmic bias (Hampton, 2021). Similarly, Braun and Hummel (2022), emphasize that the mitigation of bias neglects their social structures. The very idea of mitigating bias is misleading in the social context, because – as Birhane (2021) points out – "bias is not a deviation from the correct description" (p. 6) but represents the inequal reality. As we will see later again, the assumption of a technical fix might be harmful, because minor mathematical changes towards debiasing might legitimatize the use of an algorithm (Kasirzadeh, 2022), overlooking its actual societal impact.

The assumption of ‘fixing’ injustice also indicates an orientation towards techno-solutionism, as criticized by several authors (Andrus & Villeneuve, 2022; Gangadharan & Niklas, 2019; Green, 2022; Hampton, 2021; Heilinger, 2022). The technic-centered view limits the considered problems and thus also the available solution strategies. Challenges without a clear solution strategy, structural reforms, and social inequalities are likely to be considered outside the scope of the research on algorithmic fairness (Birhane et al., 2022; Green, 2022; Zajko, 2021). Gangadharan and Niklas (2019) conclude that technology does not play the core role in discriminatory and inequitable systems, but rather supports and accompanies them. Furthermore, since AI is a powerful tool, there is a tendency to use it always without considering whether or not it is actually the right tool (Heilinger, 2022). In line with this thought, several authors request to consider whether algorithms are necessary or useful in the first place, enabling non-algorithmic solutions and the option to disregard algorithms (Gangadharan & Niklas, 2019; Green, 2022; Hampton, 2021; Heilinger, 2022; Lin & Chen, 2022; Zajko, 2021). If the algorithm is not disregarded completely, Zimmermann and Lee-Stronach (2021) remember to consider the full range of available options: An algorithmic decision usually does not indicate any uncertainty when facing a boundary decision, as it is designed to give a prediction, e.g. whether a loan is likely to be repaid: yes or no. However, there might be other possible actions between approving the loan or denying it completely, such as giving a smaller amount or changing some conditions.

Concluding from the critical analysis of algorithmic fairness, several authors illustrate the urgent need for a reform of AI ethics by a call for change in thinking, demanding that the scope of algorithmic justice has to be widened to account for power hierarchies, unjust background conditions, and structural injustices (Cinnamon, 2017; Gangadharan & Niklas, 2019; Kasirzadeh, 2022; Kassam & Marino, 2022; Lin & Chen, 2022; Zhang, 2022). Gwagwa et al. (2022) state that a “Reflective turn in the ethics of technology is necessary” (p. 6). Zajko (2021) argues for a radical approach (in contrast to conservative approaches) to enable radical change. Birhane et al. (2022) propose a “fundamental shift—from rational to relational—in thinking” (p. 1). Gangadharan and Niklas (2019) request a “reflexive turn that decenters data and data-driven technologies in the debate on discrimination” (p. 885). Hampton (2021) demands to “reimagine technology” (p. 10) and similarly Lu et al. (2022) call for “a fundamental re-imagining of how we configure our machines” (p. 1014). But what exactly do these rethinking processes entail? This will be the subject of the following analysis of topics discussed within the literature on relational algorithmic justice.

3.3.2 Critical and Constructive Approaches

This chapter investigates the critical and constructive approaches discussed in the literature. By in-depth literature analysis, six main topics were identified: (1) argumentations to focus on disparate impact instead of intention and enable affirmative action with a focus on forward-looking responsibility, (2) critique on the categorization and measurement which are the foundation of statistical fairness, (3) discussions of the interplay between algorithms, power and capitalism, (4) the investigation of epistemic challenges, (5) elaborations of the algorithms on human solidarity, and (6) approaches to enable democratic and participatory algorithms and also their criticism. *Table 4* provides a brief overview of this evaluation. Furthermore, each topic will be described in detail in the following sections.

Table 4 Overview of the Literature Analysis: Topics, Criticism, and Constructive Approaches

Topic	Criticisms	Constructive Approaches
(1) Intentions, Disparate Impact and Responsibility	Unsatisfactory focus on intentional discrimination; Structural injustices are not considered, leading to disparate impact	Affirmative algorithms; Forward-looking responsibility; Holistic analysis of upstream and downstream effects
(2) Categorization and Measurements of Humans	Subjectivity of protected attributes; Misrepresentation leads to stigmatization; Fluid matters are conceptualized as discrete categories; Intersectionality poorly addressed; Illusory ‘solutions’	Conceptualize the selection of protected attributes as a political task; Ethical data curation with emphasis on the dynamic of knowledge; Develop relational representations of human identity; Engage with intersections of oppression
(3) Algorithms, Power and Capitalism	Algorithms extend human power; Risk of Surveillance; Reinforcement of privileges; Hidden cost of algorithms; Treatment of private data	More diversity in teams; Free choice (not) to engage with technology; Ensure control over own data
(4) Epistemic Challenges	Western Centralized Data Science; Unilateral perspectives; Constructed categories as ‘facts’; Societal norms are hard-coded	Transnational feminism and feminist theories of epistemology; Conceptualize knowledge as value-laden
(5) Social Solidarity	Increased separation; Individuals reduced to statistics; Absence of empathy	Active commitment towards social solidarity to support justice; Focus on marginalized perspectives
(6) Democratic and Participatory Algorithms	Abstract ideas, goals, or methods; Misuse and ‘participation-washing’	Participation to identify needs, enable self-determination, diversify knowledge, create shared responsibility; Methods: Design Justice, Accountability for reasonableness framework

3.3.2.1 Intentions, Disparate Impact and Responsibility

The previous section illustrated the claim that a structural perspective on algorithmic justice is essential. This requires a differentiated view on discrimination, resulting in the recommendation to develop affirmative algorithms and to focus on forward-looking responsibility, as explained below.

Discrimination Without Intention and Disparate Impact

As mentioned in *Section 2.2*, structural injustices arise from the aggregation of non-blameworthy actions. Hence, structural discrimination may exist without intention. However, the intention is relevant to the current anti-discrimination law (in the U.S. but also in Germany), as the basic idea is to avoid disparate treatment and thus intentional discrimination (Kassam & Marino, 2022). Following this idea is a flaw when facing algorithmic discrimination, because correlations inherent in the data may lead to proxy discrimination which happens without any intention or even without being noticed (Kassam & Marino, 2022). Thus, the legal focus on intentional discrimination is not suitable for the algorithmic context. The need for a structural perspective is thus especially significant (Kassam & Marino, 2022) because otherwise, it may contribute to and further increase structural injustices. As Kong summarizes: "Most engineers do not deliberately discriminate against Black women. The problem is rather that they are simply doing their jobs but their actions contribute to reproducing oppression" (p. 488).

As Hampton (2021) highlights, the "question of intention is less important than the actual outcomes upon people's lives." (p. 4). Given that algorithms act at a systematic and widespread scale, the algorithms might even be more harmful than traditional problems of discrimination (Kassam & Marino, 2022). This high scale might push harmful feedback loops between humans and machines (Lu et al., 2022). Lin and Chen (2022) further highlight that the mutual effect between AI and structural injustices is bidirectional, reinforcing oppressive structures. In this way, feedback loops may increase social inequality (Aizenberg & van den Hoven, 2020) and establish a self-fulfilling prophecy of disadvantages and inequalities (Burrell & Fourcade, 2021).

As illustrated in the background, the current statistical fairness metrics aim for some kind of parity (in terms of true positives, etc.) between subgroups. Considering the disparate impact for different individuals, their different social starting points, and the structural nature of discrimination, several authors argue the focus on fairness as statistical parity is too limited, especially in the algorithmic context (Andrus & Villeneuve, 2022; Green, 2022; Kassam & Marino, 2022; Kong, 2022; Lin & Chen, 2022; Zhang, 2022; Zimmermann & Lee-Stronach,

2021). As structural injustices may not be adequately addressed by statistical parity, algorithmic justice should focus on disparate *impact* instead of *treatment*[7]. Zimmermann and Lee-Stronach (2021) conclude that algorithmic systems should actively control structural injustices. Several other authors draw the related conclusion, that affirmative action for algorithmic justice might be justifiable, as will be elaborated briefly in the following.

Affirmative Algorithmic Justice

In light of structural injustices, several authors conclude that the use of fairness metrics in favor of marginalized groups might be justifiable, giving a rationale to prioritize certain groups. Kong (2022) argues, that "active interventions that allow a higher probability of preferable outcomes for marginalized subgroups than for privileged subgroups" (p. 491) are needed in the context of ADM to reduce the effects of structural oppression. In a similar vein, Kassam and Marino (2022) argue based on collective responsibility for different risk thresholds and weighing criteria in favor of certain groups to reduce racism. Zhang (2022) demands to consider the basic capabilities that need to be secured to relate to each other as equals. In reference to E. Anderson (1999), these capabilities correspond to three roles people take on: as humans, as participants in a system of social cooperation, and as citizens. As algorithmic decision-making systems alter the access to these basic capabilities, this may be a rationale for using different risk thresholds and affirmative algorithms (ibid). More specifically, Zhang (2022) devotes her argument to lower risk scores for Black defendants in the context of predictive sentencing. She argues that a just algorithm "must be one that takes into account the disproportionate harm of pretrial detention on Black communities' access to basic capabilities" (Zhang, 2022, p. 503). Even if some might say that unequal treatment is procedurally unjust, Zhang (2022) counters this critique by pointing out that the orientation for substantive justice might reason different treatment. In conclusion, she argues that it is the logical consequence to benefit those currently disadvantaged to avoid mirroring the current privileges and injustices. While not directly referring to affirmative algorithms, Braun and Hummel (2022) also agree that a "prioritization of those who have remained insufficiently considered up to now" (p. 6) is needed.

[7] There is actually an algorithmic fairness metric called 'Disparate Impact', aiming to represent the legal notion of disparate impact and requires statistical parity of positive outcomes across groups (Pessach and Shmueli (2022)). The legal notion focuses on practices that appear to be neutral but lead to different results for a protected group (Kassam and Marino (2022)). However, in the context of structural injustices, the notion of disparate impact does not only focus on a specific practice but includes the entire social environment, e.g., different social starting points and the different severity of consequences. A more nuanced discussion, of how far the specific fairness metric 'disparate impact' matches the critique mentioned above might be valuable.

While acknowledging the rationale for affirmative action, the most critical point is highlighted by Green (2022). He stresses the dilemma that on the one hand treating everyone equally might reproduce inequality but on the other hand, special treatment might stigmatize the respective groups. In this way, affirmative treatment might also fail to enable greater equality. Green (2022) concludes that a structural and relational response is needed, which is sketched in the following.

Structural Responses and Forward-Looking Responsibility

Based on the limited scope of mathematical fairness approaches, Green (2022) demands a methodical reform that shifts away from abstract formalizations towards substantive evaluations. Therefore, he suggests a twofold strategy that includes a relational response, meaning that upstream influences of social inequalities and hierarchies are proactively considered, as well as a structural response which focuses on the downstream effects that negative algorithmic predications cause with the aim to reduce the harmful impact and reinforcing process of algorithms on social inequalities. This might be realized by a three-step strategy (ibid.): First of all, relevant inequalities have to be identified. From this, potential reforms that act upstream and downstream should be developed. The final step is to consider the role of algorithms themselves – including the possible conclusion that the algorithm might not be the appropriate tool for reform. While this links to the previously introduced criticism about techno-solutionism, Green (2022) also highlights that "this analysis will also reveal new, fruitful roles for algorithms to complement broader efforts to combat oppression" (p. 18).

Besides this approach, numerous authors (Braun & Hummel, 2022; Kasirzadeh, 2022; Kong, 2022; Lin & Chen, 2022; Rafanelli, 2022) ground their work on the theory of Iris M. Young to highlight the interplay between algorithms and structural injustices. Especially, Lin and Chen (2022) highlight the social connection model – an approach of forward-looking responsibility – by Iris Young (2006). Young argued that the liability model of responsibility, which usually requires a control condition and a knowledge condition, is unsuitable when assigning responsibility for structural injustice. Agreeing with this observation, Lin and Chen (2022) argue that "a broader group of people beyond software engineers are responsible for pursuing AI fairness and should join collective action to shape the social structure" (p. 3). To turn collective responsibility into practice, it is highlighted that different actors have different capabilities and roles to reduce structural injustices. The stress that the four modes of reasoning as introduced by Young (2006) – namely power, privilege, interest vested in the structure, and collective ability – have to be considered to incorporate the focus on collective responsibility and power relations into approaches of algorithmic justice. This point is further continued by

Kasirzadeh (2022) highlighting some practical consequences for algorithmic justice research. Her analysis reveals the need for careful evaluations of the goals when developing algorithms for a social context, the focus on the interplay between algorithmic predictions and social power, and the attention to non-technological trade-offs with increasing interdisciplinary analyses. Also, Lin and Chen (2022) develop practical approaches according to each stage of algorithm development. A selection of these approaches includes the stage of problem selection, where risks and benefits have to be identified, the stage of data curation, focusing on the evaluation of biases, the stage of model development and validation in which associated factors and proxies of measurements have to be critically discussed, including an understanding of choices as value-laden, as well as the stage model deployment and monitoring in which the real-word impact has to be constantly evaluated.

3.3.2.2 Categorization and Measurements of Humans

The previous section highlighted criticism of algorithmic fairness metrics as a form of parity. However, the analysis of the literature also reveals criticism of the attempts and underlying assumptions when measuring algorithmic fairness in the first place, as illustrated in the following.

Ill-Definition of Protected Attributes

As introduced previously, fairness metrics are orientated towards a list of protected attributes such as gender or race. However, the definition of these protected attributes is not straightforward. Kong (2022) highlights that "there is an endless list of potential 'protected attributes'" (p. 489). The current focus is limited to the legally protected and statistically observable attributes while certain types of oppression remain excluded (ibid.). Thus, Kong (2022) concludes that defining relevant protected attributes is a political matter and not only a statistical problem. Questioning the legal scope, Aizenberg and van den Hoven (2020) also note that algorithmic decision-making may enable wrongful discrimination that is not directly related to known protected grounds.

A particular challenge to work with protected attributes is raised by unobserved characteristics. Those are characteristics such as sexual orientation and gender identity and are simply not known or available within the dataset (Tomasev et al., 2021). Without being observed, these attributes also can hardly be protected. However, it is similarly important to secure privacy, especially for vulnerable groups because they may be at risk due to involuntary information disclosure – for example, the disclosure of homosexual orientation, which is criminalized by

several countries (Tomasev et al., 2021). Andrus and Villeneuve (2022) agree that collecting more demographic data might be harmful instead of promoting fairness metrics adequately.

A further unobserved characteristic is the attribute 'class' which faces similar challenges as gender identity (Tomasev et al., 2021). In the research of algorithmic justice, social class is usually not considered. This is highlighted by Zajko (2021) who states that " 'Class bias' is a term that is more likely to appear in sociology (see Goetz 1997) than AI ethics" (p. 1051). However, the social class preservers privileges across generations, and any AI is built onto an unequal ground truth that intersects with other inequalities like gender and race (ibid.).

Misrepresentation and Oversimplification

The process of determining which characteristics need to be protected and which do not can feed directly into misrepresentation and stigmatization. Zimmermann and Lee-Stronach (2021) as well as Andrus and Villeneuve (2022) both stress that this oversimplification may introduce a misconception of who is representative for which group and counts as a similar case. This leads to reduced autonomy over personal identification and the algorithms treat individuals according to (assumed) best-fitting category (Andrus & Villeneuve, 2022). For example, when a racial group is assigned based on the observable characteristic of 'skin color' it probably doesn't match the individual self-identification to a group, which is defined based on cultural traditions and the social context.

Simultaneously the process of categorization may reinforce "oppressive or overly prescriptive categories" (Andrus & Villeneuve, 2022, p. 1715). This enables the risk of blindly relying on algorithmic categories when determining treatment differences between groups (ibid.). It can lead to the establishment of separating categories while treating them in isolation and without considering their substantive meaning. Consequently, the categorization of humans or groups can result in their misrepresentation (ibid.). This is especially concerning because the categories themselves are stigmatizing, associating individuals and groups with inferiority (Krupiy, 2020). Thus, Krupiy (2020) cautions that "the use of AI decision-making processes is likely to result in stigmatization of groups and in exacerbating unjust processes of differentiation" (p. 20)

While a limited list of protected attributes may fall short or introduce harmful misrepresentation, it might be argued that the whole practice of categorizing humans is inherently flawed because fluid subjects, such as gender or race identity, cannot be squeezed into discrete categories to be computable. For example, any distributive fairness measurement regarding gender has to quantify how to measure genders beforehand. Many fairness metrics rely on a binary and oversimplified representation of human identity, ignoring that for example,

gender identity is in fact highly fluid, depends on the social context, and may change over time (Andrus & Villeneuve, 2022; Tomasev et al., 2021). As Tomasev et al. (2021) highlight, this assumption fails to account for non-binary identities and trans people. They conclude, that categorizing fluid constructs is insufficient by definition because they are simply not measurable. In a similar vein, Andrus and Villeneuve (2022) describe this methodical misguidance as a "mismatch between efforts to make these categories legible to computers and the actual, multidimensional, and often fluid nature of class-membership" which can "undermine work around fairness from the start" (p. 1710). Leavy et al. (2021) also stress the need to understand the complex and fluid nature of the construct of race, instead of reducing it to fixed observable characteristics. To overcome this, they recommend four principles for ethical data curation. This includes (1) the understanding of underlying perspectives mapped in the data, (2) the conceptualization of knowledge as choice-based and (3) value-laden (see also *3.3.2.4 Epistemic Challenges*), and (4) the inclusion of various forms of knowledge (Leavy et al., 2021).

Another oversimplification is the representation of human identity. According to Lu et al. (2022), the concept of identity is far more complex and dynamic than is often acknowledged in AI research. Instead, categories are often mutually exclusive, which does not match reality. Furthermore, they highlight that this assumption considers identity as fixed across context and assumes an inherent ground truth, neglecting the fact that this may not represent the complex nature of human identity (Lu et al., 2022). They stress that understanding it as a system of relational processes is crucial and develop two technical approaches based on multilevel optimization and relational learning (for technical details see Lu et al., 2022).

Finally, Krupiy (2020) argues that relying solely on algorithmic processes to analyze data provides an incomplete picture of individuals, hindering the recognition of their unique talents and characteristics. They stress that this procedure of dividing people into categories can harm the protection of human diversity and reinforce harmful systems of classification and subordination.

Intersectionality Poorly Addressed

A further challenge to categorize and measure social biases is the intersectional nature of human attributes. Within the algorithmic fairness research, the dominant approach to address intersectionality considers intersectional subgroups (e.g., black women) and calculates if the statistical parity among these subgroups differs (Kong, 2022). However, this assumes that it is possible to symmetrically weight disadvantages between these subgroups. Krupiy (2020) as

well as Zimmermann and Lee-Stronach (2021) emphasize that this assumption is an error because intersecting disadvantages mutually intensify each other and may not be simply summed up. As Krupiy (2020) puts it: “The cumulative disadvantage individuals experience is greater than the sum of the factors giving rise to the disadvantage” (p 16.). Furthermore, Zimmermann and Lee-Stronach (2021) intensify the previously mentioned critique on protected attributes, as this error makes it impossible for any list of protected attributes to adequately address intersectionality.

This is also related to the critique by Kong (2022) who describes that algorithmic intersectional fairness focuses on attributes such as race and gender but may not appropriately address the intersection of oppression. Thus, intersectionality is not analyzed in its relationship to power, although this is the key element of why intersectional discrimination exists (ibid.). Furthermore, Kong (2022) recalls the notion of ‘fairness gerrymandering’ as coined by Michael Kearns et al. (2018). It refers to the situations where an algorithm meets fairness standards for each group but is still unfair to intersectional subgroups. This introduces a conceptual dilemma: either the developers have to split the population into an endless number of subgroups or they have to stop this split somewhere, leading to the arbitrary selection of protected groups (Kong, 2022).

To address these problems, Kong (2022) demands a shift of “the focus of fairness research from intersections of protected attributes to intersections of structural oppression” (p. 492) and extend the algorithmic fairness research with non-distributive approaches. This is only possible when algorithms are proactively used to challenge oppression and focus on the purpose of the algorithm instead of making it mathematically ‘fair’ in terms of unbiased (ibid.). Finally, Kong (2022) proposes collaborative and interdisciplinary projects to approach intersectionality.

Counterproductive, Potentially Misused, and Illusory Fairness Metrics

The previous criticism led to the concern that the current approaches to algorithmic fairness are counterproductive and susceptible to abuse. Birhane (2021) as well as Zimmermann and Lee-Stronach (2021) express their concern that algorithms only pretend to be just, working under a “veneer of objectivity” (Birhane, 2021, p. 2; Zimmermann & Lee-Stronach, 2021, p. 26). Burrell and Fourcade (2021) argue that the technical fairness approaches are an important component of upholding the unequal power hierarchies between those who control the technology and those who are affected by it. This deception of society is working because of the following logic: “By seemingly eliminating human arbitrariness, the new regime of algorithmic classification lends itself to an argument for procedural fairness” (Burrell & Fourcade, 2021, p. 223). Krupiy (2020) evaluates the use of algorithmic decision-making itself

as harmful because it reinforces "a hierarchical system of classification and by shifting attention away from how the construction of societal relationships disadvantages individuals" (Krupiy, 2020, p. 15).

That the proposed fairness metrics may distract attention from more urgent points of justice finds further support in the literature: Andrus and Villeneuve (2022) point out that instead of challenging structural injustices, the focus is narrowed down to the dimension of biases. Kong (2022) emphasizes that much energy is wasted on finding the right combination of identity categories or characteristics rather than addressing intersectional systems of oppression. (Kong, 2022). Following Green (2022), the effort to develop mathematical fairness metrics is counterproductive because it "obscures pathways for systematic reform" (Green, 2022, p. 90).

3.3.2.3 The Interplay of Algorithms, Power, and Capitalism

The previously described concern regarding the misuse of mathematical fairness metrics already gave some hints at the impact of algorithms on the dynamics of power and privilege. In the following, I will illustrate in more detail, which claims and concerns referring to the algorithmic influence on power dynamics are discussed within the reviewed literature and how this relates to the underlying system of capitalism.

Algorithmically Transformation of Power

Rafanelli (2022) highlights that a populistic concern regarding AI is often presented in a way that AI might gain control over humanity and shift the power from humans away towards algorithmic systems. However, she stresses that this is a dangerous misapprehension: "AI is a tool with which humans exercise power, rather than a replacement for human power, and its use therefore raises questions of justice." (Rafanelli, 2022, p. 5). AI is thus to be seen as a tool that might be used by humans – and it is up to these humans whether they use this tool in a just way or not (ibid.)

Unfortunately, several harmful examples of how algorithmic power is used are present. Burrell and Fourcade (2021) stress that algorithms might be used for surveillance and manipulation, for example by triggering addictive behavior in online environments. Cinnamon (2017) extends this point and highlights that manipulation is also a means to gain control over individuals.

As Zajko (2021) highlights the accumulated power of algorithms is "typically used as a tool to reinforce the status quo and benefit those at the center, rather than the margins" (p. 1048). This is influenced by computer scientists who may control the ADM process by determining the underlying standards and choices (ibid.) Extending this point, Krupiy (2020) as well as Braun

and Hummel (2022) observe that data-driven processes and ADM affect and change the social relationships between individuals and institutions.

All this led to the denotation of a new social class: Cinnamon (2017) illustrates that society is divided into three ‘data-classes’ (p. 615) depending on their roles of creating, collecting, or analyzing data. This introduces a “data-class status hierarchy” (p. 615) which shapes the degree of control over individual lives. Burrell and Fourcade (2021) introduce the term ‘coding elite’ (p. 215) which refers to a group of people who strengthened their power through technical control, benefitting from marginalized or unpaid workers. The power is consolidated into those who own the code and may instruct others how to use this code (Burrell & Fourcade, 2021). Thus, the coding elite operates within and benefits from a capitalistic system. In the following, I will briefly describe the discussion on how AI is developed and deployed within this background.

Exploitation, Privileges, and Economic Influences

To assess the justice of an algorithm, the system in which it is developed and used must also be considered. First of all, several authors highlight the bad working conditions and dominating oppression in the technological labor market. It takes a lot of environmental resources and invisible work is needed to make AI work (Burrell & Fourcade, 2021; Heilinger, 2022). This illusion of automation depends on the labor of crowd-workers (Burrell & Fourcade, 2021), which usually is poorly recognized and highly underpaid (Burrell & Fourcade, 2021; Kong, 2022). The poor conditions of the crowd workers are intensified due to the hierarchical and colonial labor mark, where the workers risk the job they depend on if they don’t follow the instructions (Kong, 2022). Burrell and Fourcade (2021) further highlight that the online work environment individualizes and isolates the workers. This may hinder collective solidarity (see also *Section 3.3.2.5*) as crowd-workers have no basis or means for collective mobilization. Sometimes, the needed crowd-work is hidden under the veneer of participation as Sloane et al. (2022) note. Participation and co-creation may be used to gain an economic benefit, without acknowledging or compensating the work done (ibid.). The underpaid labor contrasts with the high financial gains of the company and illustrates that power asymmetries and domination are shaping the labor market (Heilinger, 2022). Also, Burrell and Fourcade (2021) highlight that extreme wealth and poverty co-exist where technology is produced.

Another consequence of structural injustices highlighted is the fact that developers and data scientists are mostly represented by a privileged group (Huang et al., 2022; Leavy et al., 2021; Rafanelli, 2022; Tomasev et al., 2021; Zajko, 2021). Birhane (2021) specifies this group as

"mainly elite, Western, cis-gendered, and able-bodied white men" (p. 5). Tomasev et al. (2021) stress the predominating cisnormativity in the AI research. The homogenous and privileged group of developers may lead to the fact that their perspectives and experiences are built into the algorithms, developing a product that reflects those only privileged and reinforces injustices (Huang et al., 2022; Leavy et al., 2021; Rafanelli, 2022). While the first intuition to address this problem might be to increase diversity in teams, Hampton (2021) emphasizes that more diverse teams will not necessarily solve algorithmic oppression. She puts it in a nutshell describing that the focus on more diverse teams "shifts responsibility from 'our technologies are harming people' to 'BIPOC[8] tokens have to fix it'" (p. 2). Given the still-lasting forms of discrimination and power imbalances, more diversity in the teams might be necessary but not sufficient (Hampton, 2021; Tacheva, 2022).

Another interplay between capitalism and algorithmic justice is the economic incentive and influence of the companies on the research on algorithmic fairness. As Heilinger (2022) indicates, the discourse on AI Ethics itself is not only shaped by academia but also by politics and economy. He cautions that many of the AI Ethics experts might have questionable goals or conflicting interests due to their affinity with the economy. Zajko (2021) and Lu et al. (2022) highlight that AI development is situated within capitalism and is driven by the logic of capital. This dominating financial incentive may lead to questionable goals and distract the focus from justice because it is rather optimized for efficiency and not to maximize justice and may lead to conflicting strategies (Krupiy, 2020; Lu et al., 2022; Tacheva, 2022; Zajko, 2021).

Whose Data and What for?

Furthermore, the financial incentive produces challenges regarding the ownership of data. Cinnamon (2017) highlights that although data is considered a personal matter, data is traded as a material good. With an increasing separation between those who produce the data and those who control, access, and gain money from it, many exploitative agreements take effect (ibid.). With examples of the exploitation of the financial needs of homeless people and the use of images from child pornography or of dead people, Hampton (2021) reminds us of the questionable disempowering and nonconsensual data collection practices. These are also examples of the important relationship between autonomy over self-representation and privacy as Aizenberg and van den Hoven (2020) highlight: "Privacy is essential for exercising that autonomy and protection from shaming or distortion of one's personality" (p. 8). Thus, the ownership of data is interconnected with issues of privacy and (in)visibility in data.

[8] BIPOC = Black, Indigenous, and People of Color

Given that fairness metrics rely on demographic data, adherence to privacy is also a challenge for the current approaches to algorithmic fairness (Andrus & Villeneuve, 2022). The sharing of sensitive information such as demographic data might harm in particular vulnerable groups (Andrus & Villeneuve, 2022; Tomasev et al., 2021). Furthermore, Aizenberg and van den Hoven (2020) express their concern that data collection for algorithmic fairness might be misused to expand the surveillance infrastructure. While surveillance structures might lead to harmful hypervisibility of those oppressed, they are also made invisible at the same time by algorithmic errors such as flawed face recognition systems or the Twitter algorithm which highlighted white faces while cutting Blacks from the picture (Hampton, 2021). Braun and Hummel (2022) note that structural injustices made those marginalized less visible in their claims and rights and demand the ability to control engagement and thus visibility in the digital data markets by oneself.

3.3.2.4 Epistemic Challenges: Knowledge, Experiences, and 'Facts'

It is highlighted that data science is shaped by Western experience, marginalizing the knowledge of those who are considered inferior and reinforcing the perspective of those privileged (Birhane, 2021; Huang et al., 2022; Leavy et al., 2021; Tacheva, 2022). Also, prevailing research on AI ethics itself is Western-centralized (Birhane, 2021; Gwagwa et al., 2022). Consequently, several authors invoke feminist theories of knowledge and epistemology (Birhane, 2021; Huang et al., 2022; Leavy et al., 2021). Gwagwa et al. (2022) highlight that only a few unified African approaches to global AI ethics exist, indicating a lack of inclusion of diverse countries from the global south. Furthermore, Tacheva (2022) observes that approaches like data feminism are also not devoid of Western orientation, and argues for a transnational feminist approach to widen the scope of discoverable solutions (ibid.).

The present epistemic injustices are maintained due to the underlying assumption that science represents an absolute, objective, and value-free truth (Hampton, 2021; Krupiy, 2020; Tacheva, 2022). Furthermore, (western) science relies on a rational view, i.e., it appreciates rationality above emotions (Birhane, 2021; Krupiy, 2020). This view conceptualizes knowledge as stable, logical, and shaped by certainty matter (Birhane, 2021). However, the rational worldview is challenged by the emphasis on the feminist philosophy of knowledge, highlighting that scientists are subjective in their choices and assumptions, which in turn affects the conducted research (Birhane et al., 2022; Krupiy, 2020). As a path forward, Gwagwa et al. (2022) and Birhane (2021) recommend incorporating non-Western views into the algorithmic justice research. They highlight the theory of Afro-feminism, which integrates the concept of wisdom as a kind of knowledge grounded in concrete lived experience, and the African philosophy of

Ubuntu, a relational approach focusing on mutual treatment and communal relations. These feminist views stress the need to prioritize the experiences of those marginalized and affected by (algorithmic) injustice, a claim which is supported by many authors in the presented literature review (Birhane, 2021; Braun & Hummel, 2022; Gangadharan & Niklas, 2019; Hampton, 2021; Huang et al., 2022; Krupiy, 2020; Leavy et al., 2021). This claim further highlights the right of those affected to speak for themselves (Hampton, 2021) and the need to focus on experiences to detect injustice (Braun & Hummel, 2022) and bias (Huang et al., 2022).

The dominating perspectives in computer science lead to the flaw that the social construction of categories becomes forgotten: Krupiy (2020), Andrus and Villeneuve (2022), Zimmermann and Lee-Stronach (2021), and Leavy et al. (2021) point out that concepts such as gender or race have embedded social values and an ascribed social meaning which is ignored when they are reduced to calculable categories. In turn, these categories are treated as a natural and objective fact (Green, 2022; Leavy et al., 2021), leading to misconceptions as highlighted by Rafanelli (2022): "A society that adopted this system [*sexist hiring software*] might begin to see this equation of being 'qualified' with being 'like previously successful men' as an objective truth" (p. 4). It is highlighted that by relying on these epistemic misassumptions, ADM further maintains the unequal status quo, as it "naturalizes group differences in risk that are the product of oppression" (Green, 2022, p. 12) and "perpetuates a current system of differentiation" (Krupiy, 2020, p. 18). Used in social contexts such as education or employment this strengthens a fixed social order (ibid.).

All this sums up the fact that algorithmic decision-making influences human behavior and prescribes societal norms. For example, Lu et al. (2022) argue that the fixed, non-relational categories immobilize concepts of human identity and codify the existing norms. Krupiy (2020) makes this point by highlighting that algorithmic decisions entrench social norms by blindly relying on socially constructed measurements such as intelligence tests and grades. The resulting prediction "reproduces a classification system that creates hierarchies between groups" (p. 19). In this way, individuals might be increasingly forced to adapt their behavior toward the algorithmic evaluations as illustrated by the following quote:

> "There is a danger that the operation of AI decision-making processes will act as a divisive force. As a result, there will be a deepening of segregation. Individuals will adapt their behavior in order to increase their chances of receiving a positive decision from the AI decision-making process." (Krupiy, 2020, p. 17)

The authors illustrate the urgency to challenge the foundations of scientific knowledge and assumptions that are embedded into technological development. They highlight the value of human experience which also relates to the next topic: the negative influence of ADM on social solidarity based on the absence of valued experiences.

3.3.2.5 Social Solidarity and value of human decisions

Algorithmic decision-making may hinder social solidarity. This is illustrated by Krupiy (2020) by pointing out that people are treated as representatives of a group and not as individuals. Consequently, the individual becomes less likely to engage in activities typical for another group which increases separation (ibid.). This in turn can lead to the fact that solidarity between different groups is decreased. The separation may become visible in the form of constrained choices which are created due to recommender systems and result in filter bubbles. As a consequence, Cinnamon (2017) argues that personal data analytics may hinder the free development of identity, as, for example, the offered choices online are narrowed through the best-fitting filter bubble.

Furthermore, algorithmic decisions that are purely based on statistical correlations cannot be empathic which is an obstacle to solidarity. This is highlighted by Aizenberg and van den Hoven (2020) stressing that statistical predictions are based on correlation rather than causation, leading to the fact that "data are used to judge people on the basis of what the algorithm says they might do/be/become, rather than what the person actually has done and intends to do" (Aizenberg & van den Hoven, 2020, 7). Unlucky circumstances may not be taken into account and a nuanced consideration of the individual cases is lost (ibid.). In a similar vein, Krupiy (2020) stresses that an algorithmic decision fails to consider the uniqueness of individuals and the diversity of their experiences. Also, Zimmermann and Lee-Stronach (2021) emphasize that algorithms are not able to give the benefit of the doubt and also fail to acknowledge uncertainty and underlying structural injustices. This leads both Krupiy (2020) and Zimmermann and Lee-Stronach (2021) to the statement that in social contexts a human decision-maker should be preferred over an algorithm because humans can have a more holistic assessment and deliberation about individual cases.

Finally, Krupiy (2020) illustrates that the use of AI decision-making processes may create an obstacle for citizens to comprehend how unjust social relationships hinder their success, leading to more self-blaming instead of challenging the AI decision-making process as an institution. As a result, injustices become more abstract and thus less combatable. Krupiy (2020) concludes

that it is essential to consider the ADM process as an institution that changes how individuals are embedded in a set of relationships with individuals and institutions.

In the assessed literature, the focus on solidarity is also discussed as a path toward more algorithmic justice. Braun and Hummel (2022) reemphasize that the commitment towards social solidarity to eliminate algorithmic injustice would be a driver for change. Applying solidarity to algorithms implies that shared practices are developed, all voices are heard and the concerns of those marginalized are focused (ibid). Based on the philosophy of Ubuntu, Gwagwa et al. (2022) also emphasize that solidarity and relational approaches are vital for AI ethics. By incorporating the value of solidarity with a focus on communal relations which are "based on generosity, hospitality, compassion, and friendliness" (Gwagwa et al., 2022, p. 2), technology would be able to capture the value of communities and relationships more faithfully (ibid.). The focus on solidarity features a conceptual overlap with the ideal of algorithmic democratization, which will be summarized in the next section.

3.3.2.6 Algorithmic Democratization and Participation

Himmelreich (2023) describes the call to democratize as a current trend and Birhane et al. (2022) state that "AI has taken a participatory turn" (p. 1). As democratic and participatory algorithms are a trending topic, many – sometimes contradictory – claims regarding algorithmic democratization and participation have been made. In the following, I will outline a brief summary of the current discourse.

Advantages of Algorithmic Democratization and Participation

Both, instrumental and instinct values are allocated to algorithmic democratization and participation. On the instrumental side, participation is highlighted to include various forms of knowledge from society (Birhane, 2021; Huang et al., 2022; Leavy et al., 2021). Furthermore, Andrus and Villeneuve (2022) state that misrepresentation might be addressed by enabling those affected to exercise direct control by defining their group identity as self-determined. Kong (2022) suggests collaborative projects between communities and researchers to create shared responsibility. Furthermore, she illustrates that it is important to test and discuss high-stakes algorithms together with diverse stakeholders before they are used. On the intrinsic side, democratization and participation are discussed as a matter of justice. Cinnamon (2017) illustrates how the current age of personal data analysis is a threat to the normative principle of parity of participation as discussed by Fraser (2008). Following Fraser, Cinnamon (2017) highlights that the three obstacles to parity of participation (maldistribution, misrecognition, and misrepresentation) are amplified by the dominating data practices. The call for participation

and democratization is strongly related to the previously described demand to focus on the experiences of those affected. Meaningful participation requires including the voices of those affected (Heilinger, 2022), letting the decision-subjects' needs guide the development of AI (Zajko, 2021), and enabling self-determination and community empowerment (Birhane, 2021). Wong (2020) argues that decisions regarding algorithmic fairness "essentially involve choices between competing values" and therefore it "should be conceptualized first and foremost as a political question and be resolved politically" (p. 226). Green (2022) extends this statement further, arguing that next to the choice of any fairness metric, "it is also necessary to democratize decisions such as how to reform discriminatory policies and whether to use algorithms at all" (Green, 2022, p. 90).

To enable more algorithmic participation, several methods are mentioned. Aizenberg and van den Hoven (2020) offer an overview of specific methods from Design for Values, Values-sensitive design, and Participatory Design. Furthermore, the concept of 'design justice' as introduced by Costanza-Chock (2020) is recommended by several authors (Birhane, 2021; Hampton, 2021; Heilinger, 2022; Leavy et al., 2021; Zajko, 2021). To realize the democratization of AI, Wong (2020) applies in reference to Daniels and Sabin (1997) the accountability for reasonableness framework. He suggests an approach that aims to be "reason-giving and reason-responding in the exchange of reasons that show respect to individuals' views and voices and recognize their differences" (Wong, 2020, p. 239). According to Wong (2020), this requires the industry and research community to use comprehensible language and to publish the current approaches and weaknesses of algorithmic fairness; to focus on the reason and perspectives of those who are harmed by the use of AI and to develop means to approach conflicting perspectives regarding the use of AI which may enable the dynamic adoption of algorithms.

Criticism

However, there are also critical voices discussing the ideas of participation and the democratization of AI. While Himmelreich (2023) and Sloane et al. (2022) agree with the general ideas discussed above, they criticize how it is implemented. Himmelreich (2023) distinguishes democratization and participation, stating that "more democracy does not mean more participation" (p. 1337). This indicates the first point of criticism: the ideals of democratic or participatory AI are abstract ideas, lacking defined goals or methods. Himmelreich (2023) states that the calls to democratize AI are without a clear substance and point to a missing distinction between participation and deliberation. Furthermore, he reminds us of the democratic boundary problem and refers to the (probably unsolvable) challenge to include *all*

members of society in the development, design, and deployment of AI. A similar point of criticism was also emphasized by Birhane et al. (2022) about the term participation, observing a "lack of clarity on what meaningful participation entails and what it is expected to do" (p. 1). Additionally, they are questioning the motivation of industry stakeholders who are promoting participation in AI and emphasizing the importance of including non-experts in the definition of participatory AI. It is also necessary to establish clear standards for participatory AI, which should include factors such as the level of reciprocity, reflexivity, empowerment, and the duration of a task (Birhane et al., 2022). The second point of criticism is that the superficial conception inherent in calls for participation and democratization of AI makes them susceptible to the risk of misuse. Sloane et al. (2022) coin the term 'participation-washing' and highlight the risk of enabling exploitative forms of community involvement. As highlighted above, companies may gain an economic benefit by the misuse of participation as work (Sloane et al., 2022). In a similar vein, Himmelreich (2023) cautions against the ideological exploitation of participation, as it might be misused to hide the centralization of power and turned into a tool against democracy. This is related to the third and last point of criticism, namely, the observation that participatory approaches often ignore and thus fail to fight power asymmetries (Birhane, 2021; Birhane et al., 2022; Sloane et al., 2022). Himmelreich (2023) supports this point by including a quote from Cate Crawford: "To suggest that we democratize AI to reduce asymmetries of power is a little like arguing for democratizing weapons manufacturing in the service of peace" (Crawford, 2021, p. 223).

3.4 Discussion

Although the presented findings are mainly dominated by a negative view of algorithmic fairness, several authors also express the hope that algorithmic systems could be used proactively to reduce structural injustices. In this spirit, Kasirzadeh (2022) states that "there is the potential that algorithmic systems can be used to repair some problematic structures and to generate better ones" (p. 355). Leavy et al. (2021) suggest that this might be reached when quantitative data is actively used to highlight and combat racism. While emphasizing the difficult challenge of using algorithms as a driver for more justice, Zajko (2021) summarizes that "there may be ways of designing new AI systems that help to shift power, as long as this is done with the participation of the people, groups, and communities that such efforts are intended to help" (p. 1048).

It might be seen as a shortcoming of the presented SLR that no clear guidance was given for the implementation of relational algorithmic justice. However, this lies in the nature of the

presented subject. While problems without clear solution strategies are assumed to be outside the scope, relational algorithmic justice must extend the focus beyond technical and clear solutions to consider complex structural dependencies. Before any holistic solution that goes beyond the distributive scope can be found, the structural problems have to be identified first. Birhane (2021) stresses that her paper "mainly offers critical examinations and reflection and not 'solutions'" (p. 1). In a similar spirit, the strength of the presented SLR lies in the compilation of underexplored problems within algorithmic justice research. This is the necessary first step before founded guidance of relational algorithmic justice can be developed. By avoiding the fallacies of techno-solutionism combined with a nuanced problem understanding of structural injustices the hopes to use algorithms to proactively reduce structural injustices might be supported. As a directed reaction to the criticism that the status quo becomes fixed and ignores structural patterns of injustices the paper 'Roles for Computing in Social Change' by Abebe et al. (2020) has to be highlighted. They argue that computational research can help reduce social problems in a diagnostic role, as a formalizer, as a rebuttal demonstrating technological boundaries, and as a synecdoche, drawing attention to long-standing problems (ibid.). It is this interdisciplinary discourse that is essential to improve the efforts of algorithmic justice research and has the potential to enable the use of algorithms for fighting structural injustices. A differentiated view is needed, in which areas AI might help and in which it is counterproductive with context-specific evaluations. This was done by Krupiy (2020) for the case of college admissions and will also be part of the discussion of relational algorithmic justice in the case study of hiring algorithms in *Chapter 4.*

However, a limitation of the presented SLR lies in the keyword selection on relational justice. Given the wide scope of the philosophical discussion, the used keywords are only some of the most important keywords that indicate some reference to a relational thought. However, this list could be extended for example with the concepts of diversity and intersectionality. To reduce the scope of the SLR, the number of search terms remained limited to a few but important keywords. While this may not cover the entire discourse on relational egalitarianism, it is a first operationalization of the concept of relational algorithmic justice.

Furthermore, due to the limited scope of the SLR two papers were excluded although they add important perspectives to the research on relational algorithmic justice. First, the paper 'Earth for AI: A Political Ecology of Data-Driven Climate Initiatives' by Nost and Colven (2022) is an important contribution to the research on climate AI, as they illustrate how neocolonial and racist power dynamics are manifested within. A holistic perspective on structural algorithmic

justice would need to include the area of climate justice as introduced by Nost and Colven (2022). The interplay between climate justice, relational justice, and algorithms deserves a deeper investigation in further research. The second topic that might not be adequately reflected within the presented SLR was brought up by Tapu and Fa‘agau (2022) in the paper ‘A New Age Indigenous Instrument Artificial Intelligence & Its Potential for (De)colonialized Data’. Colonialism and the resulting unjust power hierarchies and disrespectful treatment of those colonized are highly relevant for relational justice and further research on the interplay between colonialism and algorithms is reasonable.

A further topic that was not directly focused on in the presented SLR is the challenge of responsibility in the context of ADM. Nevertheless, it is highly relevant for a relational algorithmic justice account to engage with the interplay of power, responsibility and social impacts potentially caused by algorithms. Kasirzadeh (2022) and also Lin and Chen (2022) stressed the benefits when a forward-looking account of responsibility is incorporated into the discussion of ADM, as this would support the development of algorithms as a tool that proactively reduces structural injustices. However, responsibility in the context of algorithms is often discussed in terms of a responsibility gap (Matthias, 2004; Santoni de Sio & Mecacci, 2021), referring to the problem that human control over algorithmic predictions might be lost and thus the attribution of moral culpability in the event of harm becomes difficult. Given that the social connection model of Young (2006) starts with the same initial observation – that the attribution of control is not suitable for her context (i.e., structural injustices) – a deeper analysis of the interconnection of responsibility, relational justice, and the algorithmic responsibility gap would be valuable for further research.

In summary, the SLR highlights the need to enhance substantive algorithmic justice which benefits from a relational account of justice and has to be based on context-specific evaluations. This motivates the following analysis of relational algorithmic justice in the case study of algorithmic decision-making in hiring.

4. Relational Algorithmic Justice in Hiring

This chapter focuses on the use of algorithmic decision-making in hiring as a case study to apply the previously discussed perspective of relational algorithmic justice. I chose hiring as a case study and contextualization because of two reasons.

First, hiring decisions are powerful and value-laden. They determine who may benefit from economic opportunities, who earns which salary, and who may gain social recognition from work. Lacking the opportunity to find an appropriate job might cause difficulties in affording daily basic needs such as food, health security, or housing. Consequently, the ability to find a job serves as a proxy for income and wealth, power, and social recognition and represents therefore a relevant social capability (c.f., E. Anderson, 1999). Additionally, finding meaningful work affects self-respect and involves the satisfaction of fulfilling our capabilities (Rawls, 2008). Institutions consolidate high power by deciding who has access to these work-related privileges. Furthermore, the decisions are value-laden because there are no clear facts that identify the best applicant. Instead, it is an interplay of competencies, appearances, sympathy, and so on. In the absence of objective facts, it is particularly difficult to formalize value-laden decisions to be processable by algorithms.

Second, ADM in hiring is despite negative examples widely used. In a survey from 2018, 60% of the tech companies planned to invest in AI-powered recruiting software (Lara et al., 2018) and large companies such as Cisco, PepsiCo, or Ikea are already working with data-driven predictive tools in their hiring processes (Bogen & Rieke, 2018). With the increasing popularity of ADM in hiring the bandwidth of different tools has grown. The range goes from interactive games aiming to analyze the applicants' cognitive, social, and emotional competencies (e.g., tool Pymetrics) to an analysis of the applicants' voice, eye contact, and mood during video interviews (e.g., tool HireVue) to the automated analysis of an applicant's CVs to identify their qualifications (e.g., tool Ideal) (Bogen & Rieke, 2018). A similar tool that gained unfortunate prominence due to its gender-biased outcomes is the Amazon hiring algorithm (Vincent, 2018) which analyzes CVs and recommends the most promising applicants. The gender bias was evident in the downgrading of CVs with women-only colleges or hobbies such as women-chess clubs (Dastin, 2018). Also, more salient forms of gendered behavior influenced the results, as the algorithms learned to value certain wordings and language expressions that correlate to a masculine language (e.g., 'executed' and 'captured') and reinforced the gender bias (ibid.).

All named tools have one function in common: they systematically compare and rank the results of new applicants to past successful hires to predict whether or not the applicant will be successful too[9]. This also explains the gender bias in the Amazon hiring algorithm, because past successful hires were mostly male and the algorithm mirrored this tendency. Furthermore, these tools are used in the screening phase in which companies assess the suitability of applicants for a particular position. While algorithms in the screening phase do not make the final decision about a hire, they may send an automated rejection of low-ranked applicants (Bogen & Rieke, 2018). There are many other hiring algorithms with different modes of operations or for different hiring phases[10]. However, to be explicit about the discussed issue, my analysis is limited to hiring algorithms that relate to the screening phase and rely on patterns in past hiring data.

The technological progress and the widespread use of hiring algorithms – despite examples of biased technology and the powerful and value-laden nature of hiring decisions – underscores the urgency to consider the interplay between algorithmic hiring and justice in more detail. So far, much of the literature follows a distributive account that mainly focuses on algorithmic biases (Hunkenschroer & Luetge, 2022). However, as introduced in *Section 2.2*, the exclusively distributive understanding of justice falls short of taking relational aspects into account. This holds for discourse on algorithmic justice in the case study of hiring, too. Furthermore, Hunkenschroer and Luetge (2022) highlight that almost all contributions to hiring algorithms lack "a normative ethical analysis of AI recruiting by linking it to an established ethical lens" (p. 1001). This research gap further motivates my argumentation.

To develop my argument, I will apply the findings from *Chapter* 3 and make use of the premises introduced in *Section 2.2.3:* the unconditional appreciation of human diversity is a matter of relational justice. This demands two aspects: the recognition of individual differences, and to secure freedom of choice. In the following argumentation, I will reveal that ADM in hiring neither recognizes individual differences nor enables freedom of choice and thus fails to meet one important condition of relational equality: the appreciation of human diversity.

[9] For example, Pymetrics lets current employees play the games and builds a predictive tool by using machine learning to asses which characteristics differentiate the employer's top performances from its colleagues. To have a baseline, the employer must provide a list of its top performers to Pymetrics. Applicants with similar results as the top performers are in the next step recommended. (Bogen and Rieke (2018))

[10] For example, algorithms that use predefined qualifications instead of the comparison to previous hires (e.g. chatbot Mya) or algorithms for the attracting phase, by providing targeted job ads (e.g. LinkedIn, Google) or sending personalized messages with an invite to apply (e.g., tool Arya) (Bogen and Rieke (2018))

4.1 Lack of Recognition of Individual Differences

In the following chapter, I argue that the use of ADM in hiring fails to recognize the individual differences in its basic function – which is orientated towards the identification of patterns – ADM is at odds with the acknowledgment of human diversity. Yet, not everyone would agree with this statement. On the contrary, some argue that ADM promotes diversity by making less biased decisions compared to humans (e.g., Houser, 2019; Kappen & Naber, 2021; Miller, 2018). It is concluded, that ADM in hiring might promote equal opportunities. In the following, I will first refute this assumption, before I discuss how the basic function of ADM fails to recognize diversity. Furthermore, I illustrate how ADM in hiring creates obstacles to social solidarity which further undermines the respect for individual differences.

4.1.1 The Issue of Bias – Human *and* Machine

Due to normative incentives, law enforcement, and business advantages, companies aim to avoid bias and discrimination in their hiring processes and increase their attention towards diversity (Brennan, 2023). For example, in 2022 nearly all companies on the Fortune 500 list are committed to diversity, equity, and inclusion according to their websites (ibid.). Yet, this is a difficult task, as racial and ethnic discrimination is a widespread and international phenomenon that has barely decreased over the years (Quillian & Midtbøen, 2020). As a response, the use of ADM in hiring to support unbiased decisions has been widely discussed as an ethical opportunity (e.g., Hunkenschroer & Luetge, 2022). However, I will present two arguments against the expectation of unbiased ADM in hiring.

First, the assumption that ADM in hiring performs better – in the sense of less bias – is based on a narrative of objective and value-free decisions. However, the findings from *Section 3.3.2.4* suggest that knowledge and decisions should rather be conceptualized in an unstable and relational manner in which rationality is not always favored above emotions and experiences (e.g., Birhane, 2021). Supporting this claim, this observation is emphasized in the following for the discussion of biases in hiring, too.

Many contributions to ADM in hiring follow the dichotomous discourse between pessimism, evaluating algorithms as a harmful tool that reinforces structural discrimination, and enthusiasm, considering algorithmic hiring as an opportunity to reduce biases (Fabris et al., 2023; Köchling & Wehner, 2020). The latter interpretation follows the intuition that algorithms, in contrast to humans, appear to be free from prejudices and personal beliefs and are often associated with more objective, consistent, and fair hiring processes (Köchling & Wehner,

2020). While the optimistic view recognizes that algorithms are also biased, they stress that the use of algorithms has to be compared to biased human decision-making (Miller, 2018). The reasoning to prefer (sill biased) algorithms over humans indicates that the logical, deductive reasoning of algorithms should be favored about the emotional, unstable human decision-making. In this way, human decisions are rendered inferior to the algorithmic outcomes. It suggests that every human, no matter their individual experiences or specific training for hiring decisions they received, seems to be less suitable for the job than algorithms because humans are per se 'significantly more biased' (Miller, 2018). In a similar vein, Kappen and Naber (2021) state that "recruiters' motivation judgments turned out to be unreliable, invalid, and sometimes biased" (p. 2) and Houser directly claims that "AI is superior to human decision-making" (p. 296). Thereby, the coding class (*see 3.3.2.3*) reinforces the narrative of faulty, biased, and unreliable human decision-making to emphasize the need to use algorithmic decision-making.

Yet, this narrative ignores two points: On the one hand, interventions to reduce discrimination in hiring should not aim to fix the biased human or replace him with a slightly less biased algorithm but should rather fix the institutional hiring process itself. Once the processes are automated by seemingly more objective algorithms, it might become even harder to adjust the increasingly complex and opaque decision-making process. To consider a small example, recall that the method 'fairness through unawareness' (i.e., the idea to hide protected attributes) did not work out for algorithms. Instead, it introduced proxy discrimination against protected groups. Many proxies wouldn't be noticed by a human decision-maker. The idea of fairness through unawareness might work pretty well for humans once certain features such as picture, gender, or origin are deleted from the CV[11]. However, by focusing only on biased outcomes, the attention is distracted from structural reforms of unfair hiring processes. Furthermore, it draws a picture of recruiters consciously relying on sexism, racism, or other *isms. While that is certainly true for some recruiters, many of them are reinforcing structural injustice without any bad intentions. As highlighted in *Section 2.2.2*, structural injustices result from various non-blameworthy actions of various actors and are embedded in institutional structures (Young, 2011). Thus, it is not only the biased recruiter who is responsible for discrimination in hiring but also the institutional forces they have to work in. Instead of using algorithms just in the same institutional environment, the institutional structures should be reformed.

[11] This example of a rather easy method to reduce human biases is already common practice in several countries. At least a picture on the CV is not required everywhere. However, at least in Germany, the current norm for designing a CV contains a lot of sensitive information, including a profile picture.

On the other hand, the narrative of faulty, biased, and unreliable human decision-making ignores the value of human experiences and emotions. Considering experiences and emotions is imperative to enable humanity and solidarity. However, the exclusive focus on the comparison between human and algorithmic bias distracts the attention from the real problem, namely that systematic biases – in recruitment and elsewhere – deny equal opportunities to some individuals, exposing them to the constant threat of oppression. For example, the so-called glass ceiling effect makes it more difficult for women to be promoted in their later careers (Cotter et al., 2001). By making it more difficult for women to pursue a successful career, they are tempted to follow traditional role models, adopting structural inequalities. Social solidarity would require understanding this problem and actively supporting women on their way to promotion by removing the gendered obstacles. Yet, these structural inequalities are not centered when human and algorithmic biases in hiring are weighed against each other. It is rather discussed about the question, of how much less biased than a human the ADM has to be, to be accepted or to be considered as 'trustworthy'. Enthusiastic argumentations seem to follow the logic of 'even if the algorithm is bad, it's currently worse' to justify the use biased of algorithms. In this way, the use of ADM in hiring seemingly provides a solution but structural injustices continue to prevail. Furthermore, it remains the question of which degree of algorithmic bias is acceptable and who ought to decide this.

The *second* reason to disprove the expectation of unbiased ADM in hiring lies in the fact that the research indicates ambiguous empirical evidence as to whether this hope might become true. As said, scholars from the optimistic perspective acknowledge that ADM is biased, too. Then, there seems to be an assumed difference between human and algorithmic biases. In a nutshell, this lies in the question of whether the bias can be corrected or not. For example, Zerilli et al. (2019) distinguish between intrinsic biases, which are value-laden, connected to emotions, and difficult to discard, and extrinsic biases, which are caused solely by the input of the system (i.e., data). Given these characteristics, human biases tend to be intrinsic, and most of the algorithmic biases are extrinsic (ibid.). Algorithmic biases would be corrected as soon as the input (e.g., training data) is improved. This assumption is shared by the optimistic accounts on ADM in hiring: For example, Houser (2019) argues that AI may tackle the diversity problem in hiring when balanced and representative data sets are used to train the algorithms.

This raises the question of how to obtain unbiased data. As illustrated by D'Ignazio and Klein (2020) data is not simply *there* to be used, but is always a collection shaped by choices, assumptions, and practices. They illustrate how data collection is a form of exercising power:

It matters who collects data about whom; it matters how data is collected; and it matters for which purpose data is collected. While the optimistic view on ADM in hiring relies on unbiased data to create unbiased algorithms, D'Ignazio and Klein (2020) stress that "data are never neutral" (p. 39). Thus, the basic premise to 'fix' extrinsic, algorithmic biases with unbiased and objective data is a flaw, when this data doesn't even exist.

Supporting this statement, it is important to note that even if the data would be free from any selection bias and reflect reality accurately, this reality is inherently biased, too (e.g., Hampton (2021), Lin and Chen (2022), see also *Section 3.3.1).* Biases in datasets represent societal biases. This is especially apparent in the context of hiring: The ability to practice in a highly valued and compensated profession is associated with certain privileges – education, social class but also gender or race. For example, the technology sector is mainly dominated by elite white men (D'Ignazio & Klein, 2020). This would be transferred to the data set which appropriately represents the current truth – but certainly, also a truth that should not be reinforced in future hiring decisions because it would also reinforce structural injustices. As a response, many providers of ADM in hiring, among them the previously mentioned tools HireVue and Pymetrics, claim to actively mitigate bias (Raghavan et al., 2020). For this purpose, the fairness measurements introduced in *Section 2.3* are used (ibid.). However, as illustrated in the SLR (see *3.3.2.2*), these approaches exhibit several problems. In the following, I will briefly demonstrate two of the most relevant problems in the context of hiring, namely the reliance on protected attributes and the failure to consider intersectionality.

The fact that approaches to algorithmic fairness are based on protected attributes is particularly problematic in the context of hiring. These attributes are usually defined by legal frameworks such as by the U.S. Equal Employment Opportunity Commission (EEOC) which prohibits discrimination against race, color, religion, sex, national origin, age, disability, or genetic information (US EEOC, 2023). However, the data input for hiring algorithms (e.g., CVs, motivation letters, videos) exhibits a high degree of variety and contains lots of sensitive information. When language or hobbies are statistical predictors for gender, they may introduce proxy discrimination as illustrated previously in the case of the Amazon hiring algorithm. Thus, the bias analysis of a certain attribute is conditional to its statistical relation to a protected attribute which must not influence the decision. While this contradicts the requirement of the unconditional appreciation of human diversity (see *Section 2.2.3*) it is also questionable whether the absence of a relation to a legally defined protected attribute is a sufficient justification for this attribute to be used.

There are many imaginable robust but irrelevant patterns in hiring data sets that cannot be captured within a fixed list of protected attributes. For example, a human bias lies in the fact that tall men are considered to be self-confident and assertive (Judge & Cable, 2004). This is reflected in hiring decisions and nearly 60% of male CEOs of Fortune 500 companies in the US are taller than 182cm, even though only 15% of the US population is that height (Gladwell & Neubauer, 2005). This is a robust pattern in the data, and assuming the height is available (or possibly to be analyzed through the CV picture or application video) it is plausible that an algorithm would evaluate this attribute as a relevant predictor for employee success.

Another pattern already considered by ADM in hiring is the hobby of lacrosse, which acted as a positive predictor of career success (Gershgorn, 2018). One possible explanation for this pattern might be that lacrosse serves as a proxy for class: it is one of the most important sports for the upper-middle-class in America which consists of highly educated and salaried professionals (Hall, 2016). As highlighted by Zajko (2021) (see also *3.3.2.2)* class bias tends to be overlooked in the discourse on algorithmic justice. However, the impact of social inequality on successful hiring decisions would be highly relevant to consider. Otherwise, the use of ADM in hiring might increase the separation between socio-economic classes. Given the high variety of imaginable patterns, it is not reasonable to declare every attribute (such as height) as protected in order to prevent arbitrary decisions. On the one hand, this might lead to an endless list of protected attributes (Kong (2022), see also *3.3.2.2*), and on the other hand, the moral evaluation of which attribute should become protected and which not is a difficult task. The moral relevance can only be determined by a careful, context-based evaluation of each case.

The problem of class bias in hiring algorithms also stresses the challenge of approaching intersectionality within the frame of algorithmic fairness measures. Intersectionality is in the algorithmic fairness research commonly measured in separated subgroups (*see 3.3.2.2*) Following this logic, gender x class (e.g., low-class men vs. high-class men vs. low-class women,) would build one subgroup. But as social class intersects with protected attributes such as race, ability, and religion (D'Ignazio & Klein, 2020), all these combinations would have to be subgroups, too. Combining class with every intersecting protected attribute would quickly sum to a high number of combinations: If the eight protected attributes are separated into four social classes (lower, middle, upper-middle, and upper class) this would already create 32 subgroups. These would have to be split up further along the intersecting attributes of gender, race, and so on, illustrating the problem of ‘fairness gerrymandering’ (Kong, 2022), (*see 3.3.2.2*). Beyond the technical unfeasibility of analyzing every subgroup, the application

context of hiring also illustrates that the sum of each disadvantage does not appropriately represent intersectional disadvantages. A low-class status intersecting with sexism or racism multiplies the obstacles to finding an appropriate job, as low-class status inhibits the opportunities to gain higher education and similar work-related requirements. Once this obstacle is overcome, the individual has to fight against gender or race bias in the hiring process, which might be harder when only very few role models exist. Furthermore, the concerned individual might have less support from their social environment, simply because fewer low-class people have the job knowledge and experience to support in the hiring process. Given the social barrier to success in higher education, fewer individuals are going to apply for a higher job because they know that they don't meet the requirements or skills needed. As statistical fairness analytics are based on those who have managed to apply, any ratio between the outcomes of the considered class subgroups does not faithfully reflect equal opportunities to gain the job. Moreover, the consequence of a negative decision might be more severe for people from a lower-class status than those from a higher class. People from a higher class can likely draw on savings until they find another good job. When no savings exist, the pressure is higher to accept worse jobs and the ability to keep applying for other jobs decreases. This illustrates that harm may not evaluated symmetrically between individuals which is even more challenging if the individuals suffer from intersecting disadvantages. Due to the constraint perspective of protected attributes, individuals who experience disadvantages based on their social class, intersections of protected attributes, or other salient discriminations are rendered indivisible and fall through the pattern of the current measurements of algorithmic fairness.

Summarizing this section, it has been demonstrated that neither the call for unbiased data nor the focus on fairness metrics succeeds in defending ADM as an unbiased tool. Hence, we should reject the optimistic claim that ADM might help to support diversity in hiring by enabling better – in the sense of less biased than humans – decisions. While the comparison of whether or not ADM might be less biased than humans is not yet definite, this

Furthermore, the section illustrated that exactly this focus on the comparison of biases is misleading. It is based on a language that downgrades humans' emotional capacities to an arbitrary and fundamentally biased process of decision-making which is inferior to a logical and objective algorithm. I argued that this assumption is a flaw because it relies on the fiction that neutral and objective decisions in hiring exist and ignores the value of humanity and solidarity in sensitive decision-making processes. All in all, tension is required whether ADM is actually less biased or in any sense 'better' than human decisions.

4.1.2 Patterns in Contrast to Diversity and its Real-World Impact

I will now defend the claim that the basic function of ADM is at odds with the recognition of diversity because diversity indicates *differences*, contrasting with the mode of operation of hiring algorithms which is based on patterns, thus *similarities*. As specified above, my analysis relies on algorithms that use data from current employees to predict the success of applicants. This restriction still includes several real-world algorithms such as the previously introduced tools HireVue, Pymetircs, or Ideal.

The basic mode of operation of hiring algorithms is to analyze large data and identify regularities in the current employee characteristics, aiming to find a predictor for job success. These algorithms follow a process of induction, which means that they derive general rules from concrete examples (Barocas et al., 2023). For example, an image recognition algorithm that should identify dogs is first trained with example pictures. The pictures are analyzed for patterns in the data which constitute rules to identify all past examples of a dog. These rules are then applied to new pictures so that the algorithm identifies dogs in future cases, too. This classification (dog or no dog) is only possible due to "the *existence of patterns* that connect the outcome of interest in a population to pieces of information that we can observe" (Barocas et al., 2023, p. 44, my emphasis). While the process of induction is illustrative using the example of an image recognition algorithm, it becomes more abstract in the context of hiring algorithms. Similar to the image recognition algorithm, a hiring algorithm is first trained with large data, representing the examples of qualified applicants. A qualified applicant might be considered as someone, who was hired, thus the algorithm is trained with data from current employees. In many cases, hiring algorithms go one step beyond considering simply past hires. Instead, the employees are ranked as 'top performers', as only successful employees should be hired again. This implies, that the employer has to provide a measurement of job success (e.g., Pymetrics). As the specification of a successful employee is a value-laden and subjective task, this subjectivity will also direct the algorithmic prediction. Nevertheless, once this specification is established this will serve as the ground truth for the algorithm to identify typical characteristics of successful employees. This is analogous to the training data with pictures containing dogs in the previous example. Instead of 'dog' or 'no dog', the classification divides into 'successful' or 'not successful'. Afterward, each new applicant is compared to this metric of typical characteristics. In this way, these algorithms assume that the best future candidates would be those who are the most similar to already hired employees (Raghavan et al., 2020). This mode of operation I will call hereinafter *similarity-based hiring*.

Similarity-based hiring might lead to the issue that the algorithm might learn heuristics such as "having a good resume meant having a resume like those of previously successful men" (Rafanelli, 2022, p. 4). As discussed previously, structural injustices that influence all past decisions will also be the moderators for the ADM predictions. Remember the example of the hiring algorithm favoring the hobby of lacrosse (Gershgorn, 2018). While this example might be a proxy for the societal class, it also illustrates the risk of similarity-based hiring, as arbitrary – but frequent – correlations might become a predictor for job success. As a consequence, those behaving like the majority are likely to have more similarities to the current employees, simply because a person with less common characteristics is less likely to be represented by the data. Thus, similarity-based hiring values similar behavior more than individuality, and unique characteristics may be considered outliers in the data. This creates an obstacle to diversity in hiring because everyone who does not fall into the homogenous norm wasn't seen by the algorithm before and does not match the classification pattern.

Through extrapolating observations from the past into the future, anything new or unknown is considered worse than the conservative of doing things. This claim is supported by Barocas et al. (2023) highlighting that "Machine learning works best in a static and stable world where the past looks like the future" (p. 261). However, adjectives such as static and stable stand in contrast to the dynamic and constantly changing society, which is characterized by political events, technological progress, or environmental changes, among other things. A world that remains as it was in the past would not be desirable given the structural injustices of racism, colonialism, sexism, and so on. However, relying on machine learning algorithms to make value-laden decisions about humans might hinder structural change because it reproduces past patterns. Similarity-based hiring algorithms are no exaptation from this. For example, when society starts to change its attitudes towards trans* people and more individuals are encouraged to show their identity, this would not be reflected in the past employee data. The ADM would – following conservative rules that discriminated against trans* people – still discriminate against them. Similarity-based hiring can only be as progressive as society was at the time of data collection, lacking behind dynamic social changes.

Moreover, it is important to note that the mere observation of patterns fails to provide an appropriate rationale for a hiring decision. Causes are substituted by correlations and instead of discussing the relevance of a certain quality for the job profile, it only analyzes whether or not this quality used to be common in the past. This leads to the failure to recognize individual qualities in both cases of positive or negative predictions: In case of a rejection, the applicant's

qualities are not valued, because they have not been common for previous employees. However, in the case of a recommendation, the applicant is *only* chosen, because he reminds of other employees. By relying on frequent patterns, the individual qualities are not recognized but it is rather a lucky circumstance for the successful applicant that many previous employees had similar qualities. Further support for this claim comes from Krupiy (2020), who stresses that ADM relies on an incomplete picture of individuals, hindering the recognition of their unique talents and characteristics (see also *Section 3.3.2.2*).

A further weakness of similarity-based approaches is that an algorithm is not able to understand whether or not frequent patterns indicate oppression or unfair discrimination and fails to consider unequal social background conditions. Due to their statistical nature, these algorithms might neither distinguish between morally (in)permissible patterns nor understand which patterns have causality for job success. A pattern robustly indicating that male managers have more success than female managers is a sign of systematic (dis)advantages and structural inequalities based on gender. In contrast, a pattern that indicates that managers with good math skills have more success than those without math skills rather indicates that 'being good at math' is a relevant skill but does not indicate oppression against those who aren't good at math. Sociological knowledge and normative reasoning are required to distinguish between the moral (in)permissibility to using certain attributes for a job recommendation. To draw another example, consider the characteristic 'language expression'. Good language expressions might in fact be a relevant skill for positions with high customer contact and reasonable to be used. But considering the intersecting characteristics of individuals it becomes challenging to define the necessary level of language skills: While simple language might be a weakness for native speakers, as they are likely to have the theoretical capabilities to express themselves eloquently, the evaluation of the language skills of a refugee might differ: Even a simple language use would indicate the ability to learn fast and the motivation to make an effort. Instead of interpreting simple language skills as a weakness, it might be appropriate to acknowledge that the applicant had to overcome several obstacles when applying for the job and is likely to improve his or her language skills further. Similar stories might be built around almost all diversity dimensions, be it gender, race, origin, social class, or age. Recognizing human diversity requires an awareness of individual differences in capabilities, which are influenced by different social statuses and background conditions (see *Section 2.2.3*). However, similarity-based hiring algorithms that blindly rely on statistical patterns are not able to recognize different social starting points.

Supporting the claim that the reliance on patterns is at odds with the recognition of differences, I stressed that ADM in hiring can be conceptualized as similarity-based hiring which uses observation of the past to predict the future. It relies on patterns without accounting for the causal relationships between frequent characteristics and required job skills. In this way, the recommendations are based on frequency distributions but fail to recognize the individual qualities. As the acknowledgment and respect of another status, qualities, and capabilities is one fundamental pillar of the recognition of diversity, it has been shown that ADM in hiring in its basic function fails to recognize diversity. Another pillar of recognition is the focus on solidarity, which will be examined in more detail in the following.

4.1.3 Obstacles to Solidarity

As highlighted in *Section 2.2.3*, solidarity is one key component of social recognition and essential for the development of self-esteem (Honneth, 1995). However, as illustrated in the SLR (see *Section 3.3.2.5*) several authors claim that ADM may hinder social solidarity. In the following, I will support this claim in the context of similarity-based hiring algorithms. Solidarity can commonly distinguished into two kinds of solidarity: *among* individuals or groups and *with* individuals or groups (Sangiovanni & Viehoff, 2023). For the context of hiring, the use of ADM leads to a reduction of solidarity for both, solidarity *among* the applicants and also solidarity from a recruiter or company *with* applicants.

For the latter, it is important to note that ADM reshapes the relationship between individuals and institutions (Krupiy, 2020), i.e., between the applicant and the company. The interaction between recruiters and applicants is reduced to – in case of automatic rejection – no contact at all. Thus, the applicant is deprived of the opportunity to explain themselves. In a human-to-human interaction, it might be possible to explain something unusual in the CV within the motivation letter so that the recruiter might look past a specific limitation or put it into perspective. But, within the logic of similarity-based hiring, all experiences that do not have a relation to previous successful applicants are rendered unimportant. This creates an obstacle to solidarity, because similarity-based algorithms deny the uniqueness of individual experiences or different achievements, leading to a social devaluation (see *Section 2.2.3*).

Furthermore, ADM in hiring is not flexible enough to reflect the complex human identity or value-laden decisions. For example, the Amazon hiring algorithm calculates a scale of 5 -stars, indicating the similarity to previous successful applicants (Kodiyan, 2019). Also, the mentioned tool Ideal assigns a letter grade (A – D) to a CV (Bogen & Rieke, 2018). Neither a 5-star rank nor a latter grade could appropriately represent a complex evaluation necessary for hiring

recommendations, as there is no weighing up of different attributes. Human individuality and their unique qualities are oversimplified to a very limited scale. This scale leaves no room for uncertainty nor for considerations such as 'the applicant is weaker in domain X but has outstanding skills in Y'. As highlighted by Zimmermann and Lee-Stronach (2021) the available options of an algorithmic prediction are pre-defined (such as the 5-star rank) and fail to consider the full range of available options. Considering rejected applications with the full range of options might reveal that the applicant might be very qualified for the profession, once certain circumstances are changed. For example, an applicant with care duties might not have the ability for a full-time job. However, the possibility of home office and adjusted meeting times might change the compatibility with family, or the position might be split into two half-time jobs. This is only possible by a perspective that is based on social solidarity with those marginalized so that the motivation to change the status quo exists. But the lack of nuanced feedback creates an obstacle to social solidarity and the individual becomes hidden from the human recruiters. As introduced in *Section 2.2.3*, to be seen and visible is an important part of social recognition. This is undermined when the hiring process does not consider the human behind the application but only relies on statistical patterns and data points.

The issue of invisibility is further strengthened for marginalized people. Not only that their individuality gets reduced to group membership, but marginalized groups are also less present in the data sets. Poor representations in data sets render those oppressed invisible. This is impressively presented in the book 'Invisible Women' (Perez, 2020), showing how a lack of data collection leads to the systematical exclusion of half of the population. This has severe real-world consequences for women such as the prescription of incorrect doses of medication or the higher risk of injuries in case of a car accident (ibid.). The lack of data and attention for those marginalized is also manifested within hiring algorithms and whiting the dominant approaches of algorithmic fairness. When an algorithm relies on binary gender categories, trans* people are ignored and in this way rendered invisible or unnormal (Crawford, 2021). An instant solution might be to use three gender categories. However, there are way more fine-grid conceptions of gender than female, male, and diverse. Considering the various gender conceptions, this might end up in a similar problem of 'fairness gerrymandering' (Kong (2022), see *3.3.2.2)*: Either an endless number of subgroups have to be used for the categorization, undermining the technical advantage for algorithmic abstraction, or the level of granular classifications is stopped at an arbitrary moment. Thus, classifications always render those invisible, that do match the proposed group boundaries. This undermines the recognition of human diversity in all its intersecting dimensions.

Additionally, given the limited representation of human beings, statistical correlations cannot be empathic (Aizenberg & van den Hoven, 2020). Yet, a sense of understanding and empathy is the precondition for the attitude of solidarity. Here, it is especially important to remember that ADM in hiring is less influential for the final decision in favor of an applicant, but rather determines the automatic rejections (Bogen & Rieke, 2018). Thus, the ADM acts to the disadvantage of weaker applications. Given that 'weaker' applications *are* more likely to be influenced by structural injustices, they might have a particular justification for increased solidarity with those rejected. However, as automated rejections are never seen by any recruiter, they would not detect the potential need for solidarity – spending only attention on the more privileged applicants.

The second kind of solidarity refers to the mutual relation *among* members of the same group (Sangiovanni & Viehoff, 2023). In the case of hiring, this might for example relate to individuals under threat of the same kind of oppression in hiring (sexism, racism, …) or in the broader sense between individuals who are (or feel) unfairly treated in the hiring process. They are united due to the negative experience of discrimination. However, with ADM involved in the hiring process, it becomes more difficult to exchange these experiences. To fight discrimination and exchange experiences on it, one needs to be aware of it in the first point. However, an opaque and abstract decision-making process unsettles those affected by how to evaluate their experiences. This can result in the self-blaming of the applicants instead of challenging the algorithmic decision itself (Krupiy (2020), see *3.3.2.5*). Critically enough, self-blame undermines self-confidence, self-respect, and self-esteem, which represent the physiological value of social recognition.

In short, ADM in hiring undermines social solidarity both among and with applicants. By deciding data about whom is collected and what classifications are modeled in which detail, solidarity is denied for all those marginalized. Unequal representations (in the dataset or classifications) reinforce systems of oppression. Furthermore, the similarity-based nature of hiring algorithms leads to the fact, that individual qualities are only acknowledged in so far as they have been important in the past. All this constitutes the claim that ADM in hiring fails to recognize individual differences and may thus not appreciate human diversity. On the contrary, ADM in hiring makes the oppressed invisible and hinders structural change by relying on conservative arguments and increasing obstacles to social solidarity. The process of rendering those oppressed invisible is also a component of cultural imperialism (Young (1990b), see

Section 2.2.2). This will be the subject of the next part of my argument, the observation that ADM in hiring undermines the freedom of choice.

4.2 Constrained Freedom of Choice

As outlined in *Section 2.2.3* the second premise for appreciating human diversity is to secure freedom of choice. However, I demonstrate in the following how ADM in hiring stipulates the status quo and argue why this is an exercise of power, amplifying cultural imperialism. Kate Crawford (2021, p. 127) pointy summarizes that a "classification is an act of power". Combined with the fact that similarity-based hiring reproduces past decisions and ignores the socially constructed nature of categories (Green (2022), Leavy et al. (2021), see *Section 3.3.2.4*), this leads to significant consequences for anyone who is not conforming to the status quo (Birhane, 2021). Supporting these statements, I will in the following illustrate the power of classifications in hiring decisions.

Following the approach of similarity-based hiring, it is not questioned whether or not the past employees deserved their position. It is taken for granted, that past decisions are based on an appropriate rationale. Yet, the preferential treatment of certain persons such as white privilege and male privilege impacts the workplace. Privileges refer to unearned advantages that people are entitled to because they belong to a certain social group or have certain dimensions of their identity (McIntosh, 1992). Seeing forms of oppression such as racism not only as a disadvantage for some but also as a privilege with puts others at an advantage helps to understand why higher professions are represented by a quite homogeneous social group. This group, then, is considered to be the norm as stressed by McIntosh (1992, p. 31):

> "Whites are taught to think of their lives as morally neutral, normative, and average, and also ideal so that when we work to benefit others, this is seen as work which will allow 'them' to be more like 'us'."

This quote can be transferred startlingly well to the use of ADM in hiring. Relying on similarity-based hiring reinforces patterns of privileges systematically, as the privileged social group becomes the norm for 'successful employees' against which every new applicant has to be measured. The reinforcement of privileges further intensifies due to the environment in which ADM is developed. As revealed in the SLR (see *Section 3.3.2.3*), developers and data scientists are mostly represented by privileged groups and mainly male, western-orientated, and cisnormative. By deciding in which granularity gender or any diversity dimension is measured, it is them to decide what might be considered when the 'fairness' of the hiring algorithm is

analyzed by some proposed metrics. As the developers are statistically less likely to have a trans* identity, they make unconscious decisions and assumptions about those 'others' which cannot be conceptualized within a binary gender frame. Or they ignore the topic completely, indicating that the problem is not important enough. As illustrated by D'Ignazio and Klein (2020) collecting data reveals *who* and *what* is valued simply because it is considered important enough to receive attention. Similarly, classifying people decides who and what receives attention when searching for the 'best' applicants. Given that (mostly) privileged developers use data from (mostly) privileged employees to train a similarity-based hiring algorithm it is likely that the resulting predictions recommend mostly privileged applicants. This feedback loop ends in reinforced unequal status quo which introduces new obstacles for everyone who does not fit into the privileged social group. Using past observations to shape the future hard-codes societal norms made by the dominating group. In hiring this might catalyze a loss of diversity as I will illustrate in the following.

Because similarity-based hiring homogenizes the rationale behind hiring decisions, the frequent use of ADM in hiring enables *systematic* discrimination. This holds within a specific company and across different companies. Focusing on the within-company level, it has to be noted that a single instance (the logic of the hiring algorithm) replaces serval humans involved in the hiring decision (Hunkenschroer & Luetge, 2022). Thus, the decision to reject an applicant represents one perspective, while in traditional human settings, there might have been a discussion and deliberation between recruiters, representing different points of view. It is also the same rationale that is applied across different job positions, resulting in a more homogenous selection across the entire company. Here, it is important to bear in mind that many companies are acting on a global scale and might use the same hiring heuristics across the globe, too. This means, that when the management uses an ADM-system to organize the human resources management, the algorithm might at once be used in different countries. Here, the specification of a successful employee becomes even more challenging: Questions such as 'To whom should the applicant fit?', 'To other employees in the respective country or the 'company culture' in general?', 'How is this culture defined?' have to be carefully elaborated. Measurements of language expressions, body language, and gestures have to be adjusted to the respective culture, otherwise, the culture of the country where the management decides to use an ADM system would be imposed to a completely different context. For example, imagine a U.S.-based company that has branch offices in India. When the ADM is based on hiring data from the U.S., it implies that the Indian employees should act like their colleagues in America. It might be a bigger effort to collect country-specific data, which – again – would only be made when the

individuals and specific countries' cultures are valued (c.f. D'Ignazio & Klein, 2020). Without careful consideration, the ADM might become an invisible means to pull one culture over the other, indicating a loss of the individual company culture or the language in the branch office. To summarize, using the same ADM system across the entire company leads to an increased homogenization of the employees in their behavior and valued qualities.

Regarding the level *across* different companies Kleinberg and Raghavan (2021) coined the term "algorithmic monocultures", describing the phenomenon that many *different* decision-makers rely on the same heuristic. This refers either to the use of exactly the same algorithm or hiring tool or could be also introduced since data-sharing and open-source projects are common for the machine-learning community. Although this usually has many advantages, such as higher efficiency and research opportunities, it may also lead to the fact that ADM in high-stake decisions are based on the same components and introduce homogenized outcomes across various contexts (Bommasani et al., 2022). Given that the same rationale is used across different companies, individuals who are disadvantaged by this rationale would be rejected from every job and encounter systemic harm (ibid.). Similarly, individuals who have favored qualities and are recommended by the ADM would have the privilege of being likely to be recommended for every company to which they apply. Thus, the risk of algorithmic homogenization further increases the gap in privileges. The reflection on the risk of homogenization confirms the observation of Krupiy (2020) that "there is a danger that the operation of AI decision-making processes will act as a divisive force" (p. 17), leading to increasing separation.

So far, this chapter illustrated that ADM in hiring is a catalyst for a loss of diversity at workplaces because it homogenizes hiring decisions and reproduces privileges by focusing on similarity-based reasoning. Furthermore, due to the definition of previously hired persons as the bar against which every new employee is measured, the use of ADM in hiring establishes a fixed norm. All other groups are considered as a deviation – or in technical terms as 'outliers'. This can be conceptualized as a feature of cultural imperialism (Young (1990b), see *Section 2.2.2*): The dominant group holds work-related privileges and uses ADM as a tool to distribute hiring recommendations along their interpretation of successful employees, determining the norms and quality standards based on the past – which represent, again, their privileges.

Given that similarity to current employees increases the chance of a positive algorithmic recommendation for employment, it creates an incentive for 'all others' to adopt the behavior of previous employees. For example, while the situation of women slowly improves, a trans-women might indicate 'too much otherness' as there is no representation of trans-women in the

employment dataset. A similarity-based hiring algorithm would reduce their opportunity to gain the job unless they suppress their identity and adapt to cis-women norms. The use of a similarity-based hiring tool enables a subtle way to reinforce the worldwide prevailing orientation of cisnormativity.

However, this forced adaptation process towards previous employees does not appear to be viewed too critically within the industry. For example, as shown in *Figure 4*, business blogs recommend watching public videos of the company managers and copying their behavior in interview videos which are analyzed by HireVue (Butcher, 2016).

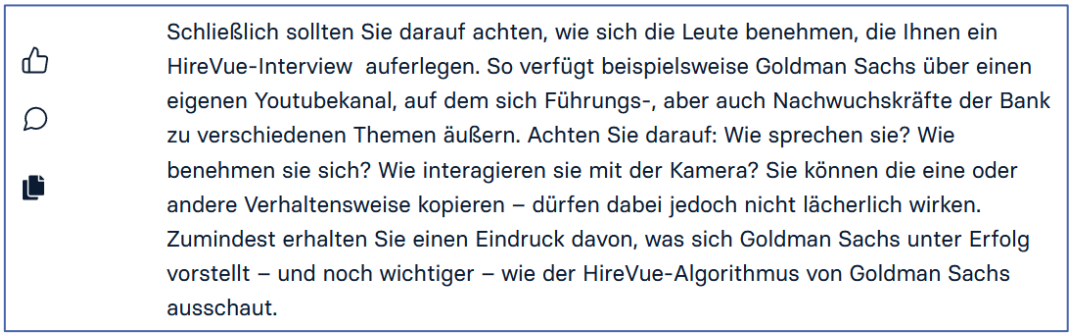

Figure 4: German blog post giving advice for job applications (Butcher, 2016)

Yet, the ability to imitate the gestures of another person indicates little meaningful reasoning for distributing job positions – at least as long as the profession sought is not an actor. The lack of a rational basis for algorithmic recommendations is referred to in the literature as *arbitrary algorithmic decisions* (Creel & Hellman, 2022). The point of arbitrariness is illustrated by the analogy from Barocas et al. (2023), describing a coach who selects his team according to the color of the runners' shoes. While this is a consistent decision-making criterion, it is arbitrary in so far as the shoe color has no relation to the goal of finding the best runners (ibid.). Similar to ADM systems, the coach uses an observed pattern without engaging with a justification of *why* this pattern might have any predictive power for his goal. Barocas et al. (2023) extend their analogy of the coach by imagining that the shoe color has indeed predictive validity for the runner's speed because the shoe color varies based on the size and runners with larger feet tend to be faster. With this analogy, Barocas et al. (2023) illustrate how arbitrary observations of patterns in the data set might become proxies for relevant information but are in themselves lacking any justification or relevance for the prediction goal. However, to discriminate based on an arbitrary characteristic by itself does indicate a form of oppression against those, who do not exhibit this specific characteristic. However considering arbitrariness together with the risk of algorithmic homogenization shows the systematic effect of ADM, which may lead to the fact that arbitrary criteria lead to a large-scale exclusion (Creel & Hellman, 2022).

While I used the fictive case by Barocas et al. (2023) to illustrate the idea of arbitrary algorithmic decisions, there are several real-world examples of arbitrary algorithmic decision-making in the context of hiring. Remember the example of the hiring algorithm which interpreted the hobby 'lacrosse' as important. The same algorithm identified the name 'Jared' as a reliable predictor for job success (Gershgorn, 2018). Or Amazon's hiring algorithm which based the recommendation on common word expressions. What these examples have in common, is that all of them have no justification for how or why it is relevant for the specific job position. It is rather a sign that some arbitrary patterns apply frequently enough among past employees to be considered meaningful. In general, it cannot be known why exactly an attribute becomes a positive predictor. It might be only based on statistical patterns or the seemingly arbitrary attribute might be a proxy for a relevant quality, such as lacrosse could be imagined as a proxy for team spirit. But companies don't care for the justification if they simply follow the ADM recommendation. Within the focus on protected attributes, such cases of arbitrary decision-making wouldn't even be noticed. Consequently, it is the ADM that influences the capabilities needed for a job.

Because the capability to find a job impacts one's status in society, this becomes a moral concern. As illustrated by E. Anderson (1999) equal opportunities to learn to play cards is not morally required because this capability does not influence the social order. However, capabilities that influence the social order are from moral worth (ibid.). By using ADM in hiring decisions, it becomes blurred which capabilities are important to consider. It increases the incentive to adapt to current employees and introduces a subtle and unconscious form of cultural imperialism. Unjustified norms determine the job opportunities or – more precisely: norms that are only justified because they are a common characteristic of the currently employed and dominant group determine the job opportunities of future applicants. In this way, the separation between the dominant group and those considered as 'others' becomes more and more fixed.

Consequently, similarity-based hiring undermines the freedom of choice because some individual characteristics become more valued than others. One might either discard the own experiences and individual diversity to adopt conservative norms or is likely to receive worse recommendations from similarity-based hiring decisions. However, not everyone has equal abilities to adapt to certain characteristics. Coming back to the analogy of the coach who chooses runners based on their shoe color, one might – once the decision-making rationale is public – use too big shoes to pretend to be in the right group and get the opportunity to prove their talent. However, too big shoes hinder fast running. Similarly, pretending to be in the

considered ‘right’ social group demands effort and creates an additional obstacle for everyone who is not part of the dominant group.

One might argue that algorithmic fairness metrics might be used to actively correct for arbitrary decision-making and thus to correct for unjustified privileges and cultural imperialism. However, the SLR revealed that the dominant fairness metrics rely on stigmatizing classifications (see *Section 3.3.2.2*). Protected attributes such as race suggest that People of Color are summarized as opponents of Whites. This neglects the fact that a Black person might share more cultural similarities with a White person than with other Blacks, depending on the social environment in which they grew up. This was taken up by the movie ‘green book’ (Farrelly, 2018) where the main character Dr. Don Shirley calls out exhausted: “So if I'm not *black* enough and if I'm not *white* enough, and not *man* enough, then tell me, Tony, what am I?”. The movie aptly illustrates how limited classifications and stigmatizing groups constrict individuals and put pressure on them to behave according to the stereotypically associated group norms. Algorithmic classifications are no exception to this. Instead, they force individuals into stereotypes shaped by the dominant group. The dominant group comes up with metrics and uses them for the justification of similarity-based hiring algorithms rather than dealing with the problems themselves. In this way, algorithmic fairness metrics also contribute to cultural imperialism, determining who and what is considered the norm and what are the ‘others’.

To summarize this section, similarity-based hiring algorithms fix arbitrary societal norms and affect the opportunities of the applicants unequally. While those who belong to the dominant groups continue to profit from their privileges, those rendered as ‘others’ have to adopt their behavior or a likely to encounter disadvantaging treatment. Consequently, ADM in hiring is a tool to exercise power, the power to maintain conservative rules and reinforce cultural imperialism inconspicuously. The applicants are facing indirect forces of oppression and are hindered in their free expression of diversity. Institutionally applied, ADM in hiring might systematically lead to homogenized outcomes – homogenized towards the dominant groups. The subordinated groups face the incentive for the rational decision to either suppress their individuality or adjust their job search to positions that historically are more common for their social group and promise a higher likelihood of being recommended by a similarity-based hiring algorithm. In this way, the freedom of choice is limited and the ADM in hiring reinforces a separation between typical jobs for those privileged and those oppressed.

4.3 Practical Implications

The previous two sections illustrated that similarity-based algorithms undermine the recognition of individual differences as well as their freedom of choice. Therefore, the use of the analyzed tools fails to realize one important condition for relational equality: the appreciation of human diversity. This observation has several practical implications for the use of ADM in hiring, which are discussed in the following.

To derive practical recommendations, it is first necessary to question what is at stake when discussing the relational algorithmic justice of hiring algorithms. One might follow the business practitioner's perspective with a focus on real facts and economic constraints. From them, the main goal would be to find a sweet spot between efficiency, legal procedures, and a general commitment to justice and diversity. This perspective shall be called *bounded* relational algorithmic justice in the following. Otherwise, one might focus on *substantive* algorithmic justice to fight structural injustices and approach a perfect and just state as closely as possible. Then, the goal is to identify an ideal social structure and societal reform to realize relational algorithmic justice in hiring without considering economic constraints. Implications for both perspectives are outlined below.

Bounded Relational Algorithmic Justice

As highlighted in *Section 3.3.2.3* the development of AI is situated within capitalism and the logic of capital and financial incentives (Lu et al., 2022; Zajko, 2021). This influence might distract the focus from justice towards the optimization for efficiency (Krupiy, 2020; Lu et al., 2022; Tacheva, 2022; Zajko, 2021). While the practitioners have to consider economic constraints, they should reconsider the use of ADM in hiring so that it at least does not increase relational injustices. To do so, several steps are important.

First of all, technical fairness metrics must be reconsidered as necessary, but not sufficient. It is a necessary condition technical fairness metrics are used to inspect the outcomes for discriminating biases. When technical fairness metrics reveal biases against gender, race, or other protected attributes, a morally critical bias has been proven and the system must not be used in the current state of development. However, this should not be confused with a sufficient condition. Even if the fairness metrics indicate bias-free outcomes within the frame of classification and protected attributes, the presented analysis of relational algorithmic justice stressed that this should not be confused with a founded statement about the justice of a system

in general. Only by reconsidering what technical fairness metrics prove and what they do not prove, the illusion of fairness as cautioned against in *Section 3.3.1* can be overcome.

Similarly, the predominating assumptions regarding the superiority of (assumed) bias-free algorithms should be challenged. Especially in the context of hiring, it is not always appropriate to rely on unemotional and statistic-based facts. Instead, the value of human experiences of both, the recruiter and applicants, should be taken into account. It is misleading to conceptualize the human recruiter per se as inferior. Instead, it should be analyzed carefully, for which specific task a human or an algorithm would be advantageous. This would differ depending on the provided information: If 100s of CVs have to be analyzed, ADM may have an advantage. However, when it comes to dynamic content such as videos or interviews, a human-to-human situation might be more appropriate for recognizing an applicant's communication skills. Especially, given that the use of new technology might cause psychological reactions such as 'technostress', i.e., "stress and concomitant psychosomatic disorder induced by the introduction of high technology" (A. Anderson, 1985, p. 6), and is likely to impact the nervosity of an applicant, a human recruiter might in creating a comfortable atmosphere in the interview.

Considering the different effects of algorithms or humans leads to the next recommendation to be generally specific about formalizations. To gain a meaningful and appropriate evaluation of the application the goals, methods, and resources have to be discussed carefully. In the example introduced *above*, it has to be questioned whether it is appropriate to search for similar-behaving persons or what are better measurements. Also, the definition of classifications such as the target variable (i.e., the measurement of successful employees) needs a founded formalization. A successful employee might be operationalized as everyone who stayed more than a specific number of years in the company (Barocas et al., 2023). The management has analyzed, who might be excluded by the number of years: For example, all recently hired employees wouldn't be represented in the data, increasing the conservativity of the decision-making rationale. Also, the new work trends suggest a higher fluctuation of jobs which might not be accounted for by a years-based target. Data from employees on parental leaves must not be excluded, and so on. The small example illustrates, that a seemingly neutral measurement such as year implies unconscious assumptions regarding a good year. Likewise, success could be measured by sales numbers, introducing distinct assumptions. Thus, it is a precondition to be specific about formalizations to assess the impacts of ADM and who might be excluded by the underlying assumption. As briefly mentioned in, the function of formalizing might be a benefit of ADM because it forces the companies to be precise about their objectives (Abebe et al., 2020). Before

any practitioner buys tools such as HireVue, it should be critically analyzed whether or not the introduced rationale and formalizations are appropriate to guide future hiring decisions

Questioning the underlying assumptions and formalizations also enables the practitioner to focus on the power of classifications. Therefore, it is necessary to always ask '*Why is a classification needed at all?*', or, '*Which value or knowledge is gained from the classification?*' and '*How is this classification justified?*'. For example, while it might make sense to classify the severity of a disease to adjust the medicine, classifications of sexual orientation to increase business profits are highly controversial (Tomasev et al., 2021). Every classification which relates directly to individual humans needs much more justification than business benefits. The same holds in the context of hiring: Classifying job-relevant skills such as math would lead to different answers to the aforementioned questions than classifying face expressions.

A very concrete recommendation for the development of hiring software is to include more flexible feedback and to enable complaints. Feedback in the form of a star rank or a letter is not satisfying to respect the other human being. Instead, the range of all options available needs to be considered (Zimmermann & Lee-Stronach, 2021). This requires overcoming one-fits-all software by involving experts (i.e., recruiters of the respective company) in the design of recommendations. Also, UX designers should be involved in questions of how to display the results to avoid unconscious nudges. For example, the order of the results might influence the recruiter, perceiving the first applicant at the top as the best one. The possibility of complaints is especially important in the interest of the applicant. The applicants should have the option to challenge the algorithmic decision. On the one hand, it ensures that the applicant is not completely at the mercy of the algorithmic decision and may escape the risk of homogenization. On the other hand, this might also benefit the company. If an applicant is as keen for a job as it reports a potential mistake, the applicant is probably confident about his or her qualities and has good arguments for the company to be considered, including a high motivation for the specific position. Thus, also from an efficiency-driven perspective, the possibility of challenging the algorithmic rejection might be beneficial.

To enable bounded rational algorithmic justice, it is furthermore essential to focus on collective, forward-looking responsibility as introduced in *Section 3.3.2.1.* Professional diversity and inclusion trainings which help to recognize individual biases and privileges should be promoted. Furthermore, these trainings should not only focus on the internal treatment of (potential) colleges, but also question the developed products. A focus on the social impacts and ethical risks is imperative for the development of technology in companies and the

education of computer scientists. This claim also receives support from the scientific literature, for example as Green (2022) emphasizes the need to include a focus on the real-world impacts of algorithms in computer science training.

Finally, practitioners aiming for bounded relational justice should consider new options for using algorithms. While the use of an algorithm only filters everyone with similar characteristics has to be challenged, it might be more fruitful to use ADM to highlight the predefined competencies of applicants. For example, such a competence could be the accurateness and diligence of an applicant. This competence can be measured based on the structure and form of the CV, writing, spelling, and so on. Based on clear criteria ADM could report this information back to both the recruiter and the applicant. Since these criteria - in contrast to lacrosse - are also relevant for the job, the applicant can work on the corresponding qualities and improve themselves. Considering the full range of how ADM in hiring might be used, it is recommended that ADM is considered as a supporting and not replacing tool. This also holds for regarding the automatization of rejections which are not to be recommended from the relational algorithmic justice perspective.

Substantive Relational Algorithmic Justice

Speaking about substantive relational justice means committing to a more just world, reducing structural injustices, and focusing on relational equality. The rethinking is also suggested by Kate Crawford (2021), from which she concludes that…

> "By asking 'why use AI?' we can question the idea that everything should be subject to the logics of statistical predictions and profit accumulation" (p. 226)

So, by asking 'why use similarity-based hiring algorithms' from a substantive relational algorithmic justice view, the previous analysis suggests dissuading from its use. It has been highlighted, that individuals facing ADM in hiring cannot explain themselves and are not recognized as individuals. Automatic rejections undermine the ability to express any form of mutual respect. Finally, similarity-based hiring is a threat to diversity, as in its basic function it observes and reproduces patterns from the past. As the past successful employees represent only a small, privileged, homogeneous group, this maintains the current status and structural injustices. The recommendation to reject ADM in hiring is not to say that the current instructional and human practices are perfectly just. Rejecting ADM in hiring in favor of substantive relational algorithmic justice also demands enabling progressive change in the hiring systems. However, this change would be even more hindered when using an ADM that

replicates conservative rules and introduces further obstacles to diversity and opaque and abstract processes. From this, three impactions can be derived.

First, to support the call from the SLR it is essential to focus on structural reform (*Section 3.3.2.1*). Rejecting ADM in favor of progressive change implies that action against structural injustices must be taken. This is less a technological question but mainly a governmental one. Remember the discussed problem of class bias in *Section 4.1.1*. Class biases can become visible in many forms within a CV. For example, many internships are unpaid which might explain a correlation to a higher social class. This reinforces work-related privileges because unpaid internships often act as an entrance to high-paid professions. Manipulating an algorithm towards a reduced class bias would likely indicate to consider internships as less important. However, as relevant internships positively impact the likelihood of job success, it might be reasoned to keep this information for the evaluation of the applicant. Instead of focusing on the algorithm, the fact that unpaid internships exist at all should be questioned. By governmental decisions, unpaid internships ought to be banned. This would remove an obstacle for people from a lower class because then everyone could afford to do an internship. (Unpaid) Internships would not be a privilege for people from a higher class anymore. In this way, structural reform changes the root causes for this small example and supports equal opportunities that go beyond an algorithmic frame. But there is also an urgent need to look directly at the algorithmic framework from a governmental perspective. The latest efforts by the German Anti-Discrimination Agency to strengthen protection against discriminatory AI (Moehl, 2024) are therefore to be encouraged.

Second, it is essential to pursue the idea of algorithmic participation and democratization as introduced in *Section 3.3.2.6.* It is a matter of relational justice to reduce the current power injustices at play. Unfortunately, a negative example from a similar application as hiring algorithms is observable in the Austrian Public Employment Services, which aims to algorithmically calculate a person's potential to reintegrate into the Austrian labor market. The algorithm is used despite its discrimination against gender and race, and – importantly – despite public protest. The protesters organized democratic exchanges in focus groups, panel discussions, and street campaigns (epicenter.works, 2024). All this resulted in proposals that clearly outline the point of view of those affected. For example, they demand that humans and not computers ought to decide about humans' fate (ibid.). However, it is not yet clear whether the protest was successful and whether its proposals are included in future decisions. Committing to democratic and participatory algorithms, this protest should be taken seriously,

or even better: not be necessary in the first place. The proactive inclusion of public discourse before the deployment of ADM is essential to secure the perspective of those affected. For hiring algorithms, this includes talking with diverse stakeholders – recruiters, newly hired employees, and potential applicants – before the management dictates the use of ADM. The same holds for the governmental decisions regarding ADM – regulations and political guidelines should be based on democratic and participatory practices, including various civil perspectives. Using democratization to reduce the stakes, educate those affected, and enable a right of co-determination or the ability to refuse the evaluation by an ADM reduces the power of the raising coding class cautioned against in *Section 3.3.2.3*. By focusing on both, the benefits of democratic and participatory AI and its current pitfalls as mentioned in *Section 3.3.2.6*, appropriate strategies to pursue the idea of algorithmic participation and democratization should be developed.

Third, the recommendation to consider new options for using algorithms is to be repeated. The focus on substantive relational algorithmic justice is not to say that all algorithms should be banned from hiring contexts. There are several good examples of how AI can be used to actively reduce injustices. Some tools prepare data by anonymizing CVs to reduce the recruiter's bias (Houser, 2019). Another example is the tool text.io, which helps to create gender-neutral job descriptions and might attract more diverse applicants. Thus, there are use cases of algorithms that might help to improve the current practices. However, these algorithms are likely to act at a different stage of the hiring phase than the selection of applicants itself and are only supporting human recruiting. Adopted from Green (2022) it has to be stressed that to identify new options for using algorithms it should always be questioned whether this usage enhances or facilitates the goal of relational algorithmic justice.

5. Conclusion

In the philosophical discourse, relational theories emerged as opponents to distributive accounts and shifted the subject of justice towards unequal power asymmetries, social relations, and structural injustices. Similarly, the 3rd wave of algorithmic justice introduces a shift in the focus from distributive and mathematical formulizations towards the conceptualization of algorithms as socio-technical systems and their impact on power asymmetries and structural injustices. Therefore, the presented thesis conceptualized the 3rd wave as *relational algorithmic justice*.

To investigate what the relational perspective adds to the – *so far merely distributive* – discourse of algorithmic justice, a systematic literature review was conducted. It crystallized the insights of 29 contributions along the underlying motivation and the discussed topics, including critical and constructive approaches. In total, six subject areas were synthesized: (1) intentions, disparate impact, and responsibility, (2) categorization and measurements of humans, (3) the interplay between algorithms, power, and capitalism, (4) epistemic challenges, (5) social solidarity, and (6) democratic and participatory algorithms. After the presentation of the findings, methodical limitations were discussed. In this context, particular emphasis should be placed on the selection of keywords on relational justice, which was not exhaustive but represents an initial operationalization. While further work is necessary to conceptualize the 3rd wave of algorithmic justice and its implementation, the presented analysis underscores the variety of topics that have received too little attention within the technical frame of algorithmic justice.

Applying the knowledge gained, implications for the case study of algorithmic-decision decision-making in hiring were discussed. Therefore, values from relational justice were put into context with current practices of similarity-based hiring. This discussion of hiring algorithms differs from the predooming comparison between human and algorithmic biases. Instead, the discussion demonstrates that it is not reasonable to abandon human decision-making, by raising the accusation of inherently inferior and biased humans to justify the use of slightly less biased algorithms. With a special emphasis that respectful social relations should be committed to the unconditional appreciation of human diversity, it becomes evident that similarity-based hiring algorithms create a barrier to diversity. Supporting this conclusion, the argumentation stressed that similarity-based hiring algorithms fail to recognize individual differences and undermine freedom of choice. Due to its mutual impacts, the discussion of cultural imperialism is especially important – and currently underexplored – when discussing algorithmic decision-making.

The normative discussion of the case study led to several recommendations to strengthen relational algorithmic justice. The main conclusion of the presented thesis is that it is imperative to incorporate relational accounts of justice into the research on algorithmic justice, striving for a more holistic understanding of justice. Especially in high-stake and value-laden use cases of algorithmic decision-making, it is essential to avoid the illusion of technical fairness based on mathematical calculations. Instead, the greater goal must always be kept in mind: Rather than searching for solutions within a restricted algorithmic framework and focusing on the algorithmically created distributions, it is more fruitful to consider the algorithmic impacts on power dynamics, social injustices, and the quality of human relations to identify new usage options of algorithms that enable structural change.

References

Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., & Robinson, D. G. (2020). Roles for Computing in Social Change. https://dl.acm.org/doi/pdf/10.1145/3351095.3372871

ACM. (2023, September 26). *About the ACM Digital Library*. https://dl.acm.org/about

Aizenberg, E., & van den Hoven, J. (2020). Designing for human rights in AI. *Big Data & Society*, *7*(2), 205395172094956. https://doi.org/10.1177/2053951720949566

Anderson, A. (1985). Technostress. Another Japanese discovery. *Nature*, *317*(6032), 6. https://doi.org/10.1038/317006b0

Anderson, E. (1999). What is the Point of Equality? *Ethics*, *109*(2), 287–337. http://www.jstor.org/stable/2989479

Andrus, M., & Villeneuve, S. (2022). Demographic-Reliant Algorithmic Fairness: Characterizing the Risks of Demographic Data Collection in the Pursuit of Fairness. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1709–1721). ACM. https://doi.org/10.1145/3531146.3533226

Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2016). Machine Bias. *Auerbach Publications*, pp. 254–264. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arneson, R. (1989). Equality and equal opportunity for welfare. *Philosophical Studies*, *56*(1), 77–93. https://doi.org/10.1007/bf00646210

Arneson, R. (2013). Egalitarianism. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2013). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/egalitarianism/

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities* (Vol. 1). MIT Press.

Barocas, S., & Selbst, A. D. (2016). *Big Datas Disparate Impact.* https://doi.org/10.15779/Z38BG31

Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. In *Conference on fairness, accountability and transparency* (pp. 149–159).

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns (New York, N.Y.)*, *2*(2). https://doi.org/10.1016/j.patter.2021.100205

Birhane, A., Isaac, W., Prabhakaran, V., Díaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022). Power to the People? Opportunities and Challenges for Participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22).* ACM. https://doi.org/10.1145/3551624.3555290

Bogen, M., & Rieke, A. (2018). Help Wanted - An Exploration of Hiring Algorithms, Equity and Bias.

Bommasani, R., Creel, K., Kumar, A., Jurafsky, D., & Liang, P. (2022). Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? *Advances in Neural Information Processing Systems*, *35*.

Borg, J. S. (2021). Four investment areas for ethical AI: Transdisciplinary opportunities to close the publication-to-practice gap. *Big Data & Society*, *8*(2), 205395172110401. https://doi.org/10.1177/20539517211040197

Branford, J. (2023). Experiencing AI and the Relational 'Turn' in AI Ethics. In *International Conference on Computer Ethics.*

Braun, M., & Hummel, P. (2022). Data justice and data solidarity. *Patterns (New York, N.Y.)*, *3*(3), 100427. https://doi.org/10.1016/j.patter.2021.100427

Brennan, J. (2023). Diversity for Justice vs. Diversity for Performance: Philosophical and Empirical Tensions. *Journal of Business Ethics*, *187*(3), 433–447. https://doi.org/10.1007/s10551-022-05278-9

Bui, M., & Noble, S. U. (2020). We're missing a moral framework of justice in artificial intelligence: on the limits, failings, and ethics of fairness. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford handbook of ethics of AI.* Oxford University Press. https://books.google.de/books?hl=de&lr=&id=8PQTEAAAQBAJ&oi=fnd&pg=PA163&ots=uD7ynk26ZE&sig=xZtw8xs6p-kegH19_e-r8Kan4XM#v=onepage&q&f=false

Burr, C., & Leslie, D. (2023). Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies. *AI and Ethics*, *3*(1), 73–98. https://doi.org/10.1007/s43681-022-00178-0

Burrell, J., & Fourcade, M. (2021). The Society of Algorithms. *Annual Review of Sociology*, *47*(1), 213–237. https://doi.org/10.1146/annurev-soc-090820-020800

Butcher, S. (2016, August 23). HireVue: Die schöne neue Welt der Vorstellungsgespräche. *EFinancialCareers*. https://www.efinancialcareers.de/nachrichten/2016/08/hirevue-die-schone-neue-welt-der-vorstellungsgesprache

Caton, S., & Haas, C. (2020). Fairness in Machine Learning: A Survey. *Cornell University*. https://arxiv.org/abs/2010.04053

Cinnamon, J. (2017). Social Injustice in Surveillance Capitalism. *Surveillance & Society*, *15*(5), 609–625.

Clarivate. (2023). *Web of Science Coverage Details*. https://clarivate.libguides.com/librarianresources/coverage

Cohen, G. A. (1989). On the Currency of Egalitarian Justice. *Ethics*, *99*(4), 906–944. https://doi.org/10.1086/293126

Corbett-Davies, S., & Goel, S. (2018, July 31). *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning.*

Costanza-Chock, S. (2020). *Design Justice: Community-Led Practices to Build the Worlds We Need*. MIT Press.

Cotter, D. A., Hermsen, J. M., Ovadia, S., & Vanneman, R. (2001). The Glass Ceiling Effect. *Social Forces*, *80*(2), 655–681. https://doi.org/10.1353/sof.2001.0091

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Creel, K., & Hellman, D. (2022). The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems. *Canadian Journal of Philosophy*, *52*(1), 26–43. https://doi.org/10.1017/can.2022.3

Crenshaw, K. (1989). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum*, 139–167.

Cudd, A. E. (2006). *Analyzing oppression*. *Studies in feminist philosophy*. Oxford University Press.

Daniels, N., & Sabin, J. (1997). Limits to Health Care: Fair Procedures, Democratic Deliberation, and the Legitimacy Problem for Insurers. *Philosophy & Public Affairs*, *26*(4), 303–350. https://doi.org/10.1111/j.1088-4963.1997.tb00082.x

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism. <Strong> ideas series*. The MIT Press.
Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference.* ACM. https://doi.org/10.1145/2090236.2090255
Dworkin, R. (1981). What is Equality? Part 2: Equality of Resources. *Philosophy & Public Affairs*, *10*(4), 283–345. https://doi.org/10.4324/9781315199795-7
Elsevier. (2023, September 17). *About Scopus - Abstract and citation database*. https://www.elsevier.com/solutions/scopus
epicenter.works. (2024, January 12). *Stoppt den AMS-Algorithmus!* epicenter.works - Plattform Grundrechtspolitik. https://amsalgorithmus.at/de/
Fabris, A., Baranowska, N., Dennis, M. J., Hacker, P., Saldivar, J., Borgesius, F. Z., & Biega, A. J. (2023, September 25). *Fairness and Bias in Algorithmic Hiring*. http://arxiv.org/pdf/2309.13933v1
Farrelly, P. (Director). (2018). *Green Book*. Eine besondere Freundschaft.
Fraser, N. (2008). Abnormal Justice. *Critical Inquiry*, *34*(3), 393–422. https://doi.org/10.1086/589478
Gangadharan, S. P., & Niklas, J. (2019). Decentering technology in discourse on discrimination. *Information, Communication & Society*, *22*(7), 882–899. https://doi.org/10.1080/1369118X.2019.1593484
Gardenswartz, L., & Rowe, A. (2003). *Diverse teams at work: Capitalizing on the power of diversity* (2nd ed.). Irwin.
Gershgorn, D. (2018). *Companies are on the hook if their hiring algorithms are biased*. https://qz.com/1427621/companies-are-on-the-hook-if-their-hiring-algorithms-are-biased
Gladwell, M., & Neubauer, J. (2005). *Blink: Die Macht des Moments*. Campus Verl.
Goldman, B., & Cropanzano, R. (2015). "Justice" and "fairness" are not the same thing. *Journal of Organizational Behavior*, *36*(2), 313–318. https://doi.org/10.1002/job.1956
Gosepath, S. (2001). *Equality*. https://plato.stanford.edu/entries/equality/
Green, B. (2018). *Putting the J(ustice) in FAT - Berkman Klein Center Collection - Medium*. https://medium.com/berkman-klein-center/putting-the-j-ustice-in-fat-28da2b8eae6d
Green, B. (2022). Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness. *Philosophy & Technology*, *35*(4). https://doi.org/10.1007/s13347-022-00584-6
Gusenbauer, M., & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods*, *11*(2), 181–217. https://doi.org/10.1002/jrsm.1378
Gwagwa, A., Kazim, E., & Hilliard, A. (2022). The role of the African value of Ubuntu in global AI inclusion discourse: A normative ethics perspective. *Patterns (New York, N.Y.)*, *3*(4), 100462. https://doi.org/10.1016/j.patter.2022.100462
Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8
Hall, T. (2016). *Social Class and Sports*. https://brokenclipboard.wordpress.com/2016/04/11/social-class-and-sports-2/
Hampton, L. M. (2021). Black Feminist Musings on Algorithmic Oppression. In *Conference on Fairness, Accountability, and Transparency (FAccT '21).* ACM.

Häußermann, J. J., & Lütge, C. (2022). Community-in-the-loop: Towards pluralistic value creation in AI, or—Why AI needs business ethics. *AI and Ethics*, *2*(2), 341–362. https://doi.org/10.1007/s43681-021-00047-2

Heilinger, J.-C. (2022). The Ethics of AI Ethics. A Constructive Critique. *Philosophy & Technology*, *35*(3). https://doi.org/10.1007/s13347-022-00557-9

Himmelreich, J. (2023). Against "Democratizing AI". *AI & SOCIETY*, *38*(4), 1333–1346. https://doi.org/10.1007/s00146-021-01357-z

HLEG. (2019). *High-Level Expert Group on Artificial Intelligence: Ethics Guidelines for Trustworthy AI.* European Commission.

Honneth, A. (1995). *The Struggle for Recognition: The Moral Grammar of Social Conflicts*. Translated by Joel Anderson. The MIT Press.

Houser, K. A. (2019). Can AI solve the diversity problem in the tech industry: Mitigating noise and bias in employment decision-making. *Stan. Tech. L. Rev.*(22), 290. https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/stantlr22§ion=9

Huang, L. T.-L., Chen, H.-Y., Lin, Y.-T., Huang, T.-R., & Hun, T.-W. (2022). Ameliorating Algorithmic Bias, or Why Explainable AI Needs Feminist Philosophy. *Feminist Philosophy Quarterly*, *8*(3/4).

Hunkenschroer, A. L., & Luetge, C. (2022). Ethics of AI-Enabled Recruiting and Selection: A Review and Research Agenda. *Journal of Business Ethics*, *178*(4), 977–1007. https://doi.org/10.1007/s10551-022-05049-6

Iser, M. (2013). Recognition. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2013). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/recognition/

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Joshua D. Harris, Carmen E. Quatman, M.M. Manring, Robert A. Siston, & and David C. Flanigan (2013). How to Write a Systematic Review. https://journals.sagepub.com/doi/pdf/10.1177/0363546513497567

Judge, T. A., & Cable, D. M. (2004). The effect of physical height on workplace success and income: Preliminary test of a theoretical model. *The Journal of Applied Psychology*, *89*(3), 428–441. https://doi.org/10.1037/0021-9010.89.3.428

Kappen, M., & Naber, M. (2021). Objective and bias-free measures of candidate motivation during job applications. *Scientific Reports*, *11*(1), 21254. https://doi.org/10.1038/s41598-021-00659-y

Kasirzadeh, A. (2022). Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy. In *AIES '22: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery (ACM)* (pp. 348–356). https://doi.org/10.1145/3514094.3534188

Kassam, A., & Marino, P. (2022). Algorithmic Racial Discrimination: A Social Impact Approach. *Feminist Philosophy Quarterly*, *8*(3/4), Article 4.

Kind, C. (2020, August 23). The term 'ethical AI' is finally starting to mean something. *VentureBeat*. https://venturebeat.com/ai/the-term-ethical-ai-is-finally-starting-to-mean-something/

Kitchenham, B. (2004). Procedures for Performing Systematic Reviews. *Keele, UK, Keele University*, *33*, 1–26.

https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=29890a936639862f45cb9a987dd599dce9759bf5
Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(22). https://doi.org/10.1073/pnas.2018340118
Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, *13*(3), 795–848. https://doi.org/10.1007/s40685-020-00134-w
Kong, Y. (2022). Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 485–494). ACM. https://doi.org/10.1145/3531146.3533114
Krupiy, T. (2020). A vulnerability analysis: Theorising the impact of artificial intelligence decision-making processes on individuals, society and human diversity from a social justice perspective. *Computer Law & Security Review*, *38*, 105429. https://doi.org/10.1016/j.clsr.2020.105429
Lara, K. de, Holden, L., & Trigg, M. (2018). Entelo's 2018 Recruiting Trends Report. https://cdn2.hubspot.net/hubfs/202646/Entelo%27s%202018%20Recruiting%20Trends%20Report.pdf?t=1530708036795
Leavy, S., Siapera, E., & O'Sullivan, B. (2021). Ethical Data Curation for AI: An Approach based on Feminist Epistemology and Critical Theories of Race. In *Conference on Artificial Intelligence, Ethics and Society (AIES)*, Virtual Event, USA.
Lee, M. S. A., Floridi, L., & Singh, J. (2021). Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. *AI and Ethics*, *1*(4), 529–544. https://doi.org/10.1007/s43681-021-00067-y
Lin, T.-A., & Chen, P.-H. C. (2022). Artificial Intelligence in a Structurally Unjust Society. *Feminist Philosophy Quarterly*, *8*(3/4), Article 3.
Lu, C., Kay, J., & McKee, K. R. (2022). Subverting machines, fluctuating identities: Re-learning human categorization. In *FAccT '22* (pp. 1005–1015). https://doi.org/10.1145/3531146.3533161
Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*, 175–183.
McIntosh, P. (1992). White privilege: Unpacking the invisible knapsack. In A. M. Filor (Ed.), *Multictlturalism* (pp. 30–36).
McQueen, P. (2023). Recognition, Social and Political |. *Internet Encyclopedia of Philosophy*. https://iep.utm.edu/recog_sp/#SH3b
Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607
Metringer, T. (2019, April 8). EU guidelines: Ethics washing made in Europe. *Der Tagesspiegel*. https://www.tagesspiegel.de/politik/ethics-washing-made-in-europe-5937028.html
Michael Kearns, Seth Neel, Aaron Roth, & Zhiwei Steven Wu (2018). Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. *International Conference on Machine Learning*, 2564–2572. http://proceedings.mlr.press/v80/kearns18a.html

Miller, A. P. (2018). Want Less-Biased Decisions? Use Algorithms. *Harvard Business Review*(26). https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, *8*(1), 141–163.

Moehl, K. (2024). *Ataman will Schutz vor diskriminierender KI stärken.* Antidiskriminierungsstelle des Bundes, beim Bundesministerium für Familie, Senioren, Frauen und Jugend. https://www.antidiskriminierungsstelle.de/SharedDocs/aktuelles/DE/2023/20240111_ki_fachgespraech.html

Munn, L. (2022). *The Uselessness of AI Ethics*. https://www.researchgate.net/publication/361151812_The_Uselessness_of_AI_Ethics

Narayanan, A. (2018, March 1). *21 fairness definitions and their politics*. https://www.youtube.com/watch?v=jIXIuYdnyyk

Nath, R. (2020). Relational egalitarianism. *Philosophy Compass*, *15*(7). https://doi.org/10.1111/phc3.12686

Nussbaum, M. C. (2003). Capabilities as fundamental entitlements. *Feminist Economics*, *9*, 33–59.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews*, *10*(1), 89. https://doi.org/10.1186/s13643-021-01626-4

Perez, C. C. (2020). *Invisible women: Exposing data bias in a world designed for men*. Random House UK. https://doi.org/10.1080/24701475.2020.1837580

Pessach, D., & Shmueli, E. (2022). A Review on Fairness in Machine Learning. *ACM Computing Surveys*, *55*(3), 1–44.

Phillips, A. (2021). *Unconditional equals*. Princeton University Press. https://moodle.rwth-aachen.de/pluginfile.php/2295177/mod_resource/content/1/1_Phillips-2021-Unconditional%20Equals.pdf

The PhilPapers Foundation. (2023, November 6). *PhilPapers: Online Research in Philosophy*. https://philpapers.org/

Quillian, L., & Midtbøen, A. H. (2020). *Comparative Perspectives on Racial Discrimination in Hiring: The Rise of Field Experiments*. https://www.duo.uio.no/bitstream/handle/10852/86230/2/WP-20-46.pdf https://doi.org/10.31235/osf.io/5z6a2

Rafanelli, L. M. (2022). Justice, injustice, and artificial intelligence: Lessons from political theory and philosophy. *Big Data & Society*. https://journals.sagepub.com/doi/pdf/10.1177/20539517221080676

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices, *104*, 469–481. https://doi.org/10.1145/3351095.3372828

Rawls, J. (1958). Justice as Fairness. *The Philosophical Review*, *67*(2), 164. https://doi.org/10.2307/2182612

Rawls, J. (1999). *A Theory of Justice: Revised Edition*. Harvard University Press. https://www.cita.or.id/wp-content/uploads/2016/06/John-Rawls-A-Theory-of-Justice-Belknap-Press-1999.pdf

Rawls, J. (2008). Political Liberalism. In *The New Social Theory Reader* (2nd Edition). Routledge.

Robert, L. P., Pierce, C., Marquis, L., Kim, S., & Alahmad, R. (2020). Designing fair AI for managing employees in organizations: a review, critique, and design agenda. *Human–Computer Interaction*, *35*(5-6), 545–575. https://doi.org/10.1080/07370024.2020.1735391

Robeyns, I., & Byskov, M. F. (2011). *The Capability Approach*. https://plato.stanford.edu/entries/capability-approach/#AckHumDiv

Sangiovanni, A., & Viehoff, J. (2023). *Solidarity in Social and Political Philosophy*. https://plato.stanford.edu/entries/solidarity/#NatuSoli

Santoni de Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology, 34*(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Scheffler, S. (2015). The Practice of Equality. In C. Fourie, F. Schuppert, & I. Wallimann-Helmer (Eds.), *Social Equality.* Oxford University Press.

Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2022). Participation Is not a Design Fix for Machine Learning. In *EAAMO '22: Equity and Access in Algorithms, Mechanisms, and Optimization*.

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research, 104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Steel, D., Fazelpour, S., Gillette, K., Crewe, B., & Burgess, M. (2018). Multiple diversity concepts and their ethical-epistemic implications. *European Journal for Philosophy of Science*, *8*(3), 761–780. https://doi.org/10.1007/s13194-018-0209-5

Sundar, S. S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. In *Proceedings of the 2019 CHI Conference on human* (pp. 1–9).

Tacheva, Z. (2022). Taking a critical look at the critical turn in data science: From "data feminism" to transnational feminist data science. *Big Data & Society, 9*(2), 205395172211129. https://doi.org/10.1177/20539517221112901

Tapu, I. F., & Fa'agau, T. K. (2022). A New Age Indigenous Instrument: Artificial Intelligence & Its Potential for (De)colonialized Data. *Harvard Civil Rights-Civil Liberties Law Review, 57*, 715 - 753.

Temkin, L. S. (1993). *Inequality*. *Oxford ethics series*. Oxford University Press.

Tomasev, N., McKee, K. R., Kay, J., & Mohamed, S. (2021). Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities. In M. Fourcade, B. Kuipers, S. Lazar, & D. Mulligan (Eds.), *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 254–265). ACM. https://doi.org/10.1145/3461702.3462540

US EEOC. (2023, February 21). *Overview*. https://www.eeoc.gov/overview

Vincent, J. (2018, October 10). Amazon reportedly scraps internal AI recruiting tool that was biased against women. *The Verge*. https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report

Voigt, K. (2007). The Harshness Objection: Is Luck Egalitarianism Too Harsh on the Victims of Option Luck? *Ethical Theory and Moral Practice*, *10*(4), 389–407. https://doi.org/10.1007/s10677-006-9060-4

Voigt, K., & Wester, G. RELATIONAL EQUALITY AND HEALTH.

Wong, P.-H. (2020). Democratizing Algorithmic Fairness. *Philosophy & Technology*, *33*(2), 225–244. https://doi.org/10.1007/s13347-019-00355-w

Young, I. M. (1990a). Displacing the Distributive Paradigm (Chapter 1). In I. M. Young (Ed.), *Justice and the politics of difference* (pp. 15–38). Princeton University Press. https://www.degruyter.com/document/doi/10.1515/9781400839902-004/html?lang=de

Young, I. M. (1990b). Five Faces of Oppression (Chapter 2). In I. M. Young (Ed.), *Justice and the politics of difference* (pp. 39–65). Princeton University Press. https://contensis.uwaterloo.ca/sites/courses-archive/1185/PHIL-324/media/documents/10a-young-1990-five-faces-of-oppression.pdf

Young, I. M. (Ed.). (1990c). *Justice and the politics of difference*. Princeton University Press.

Young, I. M. (2006). Responsibility and global justice: A social connection model. *Social Philosophy and Policy*, *23*(01), 102. https://doi.org/10.1017/s0265052506060043

Young, I. M. (2011). *Responsibility for justice*. *Oxford political philosophy*. Oxford University Press.

Zajko, M. (2021). Conservative AI and social inequality: conceptualizing alternatives to bias through social theory. *AI & SOCIETY*, *36*(3), 1047–1056. https://doi.org/10.1007/s00146-021-01153-9

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? *Philosophy & Technology*, *32*(4), 661–683. https://doi.org/10.1007/s13347-018-0330-6

Zhang, M. (2022). Affirmative Algorithms: Relational Equality as Algorithmic Fairness. In *ICPS, FAccT 2022: 2022 5th ACM Conference on Fairness, Accountability, and Transparency : June 21-24, 2022, Seoul, South Korea* (pp. 495–507). The Association for Computing Machinery. https://doi.org/10.1145/3531146.3533115

Zimmermann, A., & Lee-Stronach, C. (2021). Proceed with Caution.

Appendix A: Preprocessing

```
import pandas as pd
import re

pd.set_option('display.max_columns', None)

scopus = pd.read_csv("./data_unprocessed/scopus.csv")
scopus = scopus[['Authors', 'Title','Abstract','Year']]
scopus.head()


wos = pd.read_excel("./data_unprocessed/webofscience.xls")
wos =wos[['Authors', 'Article Title','Abstract','Publication Year']]
wos = wos.rename(columns={'Article Title': 'Title', 'Publication
Year': 'Year'})
wos.head()


philpapers = pd.read_excel("./data_unprocessed/philpapers_noD.xlsx")
philpapers =philpapers[['author', 'title', 'abstract','Year']]
philpapers = philpapers.rename(columns={'author': 'Authors',
       'title': 'Title', 'abstract': 'Abstract'})
philpapers.head()

acm = pd.read_csv("./data_unprocessed/ACM.csv")
acm = acm[['Authors', 'Title', 'Abstract','Publication year']]
acm = acm.rename(columns={'Publication year': 'Year'})
acm.head()


frames = [scopus, philpapers, wos,acm]

full_data = pd.concat(frames)
full_data.to_excel("0_final_data_concat.xlsx")


# Remove punctuation
full_data['title_processed'] = full_data['Title'].map(lambda x:
re.sub('[,\.:\-!?“”"’\'\t*]', '', x))

# Convert the titles to lowercase
full_data['title_processed'] =
full_data['title_processed'].map(lambda x: x.lower())

data_without_doublicates=full_data.drop_duplicates(subset=['title_pr
     ocessed'])
data_without_doublicates.head()

data_without_doublicates.to_excel("1_final_data_noD.xlsx")
```