universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Investigating Fairness in Recommender Systems
A Systematic Review

verfasst von | submitted by

Klara Larissa Prema Howorka BA

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 808
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Gender Studies
Betreut von | Supervisor:	Dr.techn. Katta Spiel BA B.Sc. M.Sc.

Abstract (Deutsch)

Fairness in Empfehlungssystemen ist ein Forschungsproblem, das in den letzten Jahren viel Aufmerksamkeit erregt hat, insbesondere mit der weit verbreiteten Verwendung von Recommender Systems, die auf Machine Learning basieren. Das Ziel, diese Systeme fair zu halten, hat viele Strategien zur Milderung von Biases hervorgebracht, die zur Diskriminierung von Personen auf Ebene von geschützten Merkmalen wie race oder gender führen. Diese Masterarbeit untersucht, auf welche Art und Weise Begriffe wie „Fairness“, „Bias“ und „Gender“ derzeit in der Empfehlungssystemforschung definiert werden, welche Biases am häufigsten behandelt werden und welche Arten von Strategien zur Bias-Minderung am meisten verwendet werden. Diese systematische Kategorisierung wurde mittels einer deduktiven qualitativen Inhaltsanalyse an einem Korpus bestehend aus 24 Forschungsarbeiten durchgeführt, die sich explizit mit dem Problem der Fairness und Bias-Minderung in Empfehlungssystemen befassen. Die Ergebnisse dieser Analyse zeigen, dass Fairness in diesem Bereich tendenziell als benutzerseitige Gruppengerechtigkeit definiert wird und dass die Geschlechtsvariable häufig undefiniert bleibt oder in einem binären Kontext operationalisiert wird. Darüber hinaus können die Arten von Bias, die am häufigsten behandelt werden, der Kategorie des demographic bias zugeordnet werden und stammen tendenziell aus der Modellbildungs-Pipeline. Strategien zur Bias-Minderung kommen am häufigsten während und nach der Verarbeitung durch das Modell zum Einsatz. Die Offenlegung dieser Tendenzen der Definition, Instrumentalisierung und Quantifizierung sozialer bzw. sozial konstruierter Konzepte trägt dazu bei, zu kontextualisieren, wie Wissen und Bedeutung im Bereich der Empfehlungssystemforschung produziert werden.

Abstract (English)

Fairness in recommender systems is a research problem which has gained a lot of attention in recent years, especially with the widespread use of machine learning based recommender systems. The aim to keep these systems fair has yielded many strategies to mitigate biases which result in discrimination of individuals who belong to protected attribute groups such as race or gender. This master thesis explores how terms such as ‘fairness’, ‘bias’ and ‘gender’ are currently defined in recommender system research, which biases are most commonly treated, as well as which types of bias mitigation strategies are used the most. The systematic review was conducted via a deductive Qualitative Content Analysis on a corpus consisting of 24 research papers which explicitly deal with the problem of fairness and bias mitigation in recommender systems. The results of this analysis show that fairness in this field is generally defined as user-side group fairness and that gender is usually operationalized within a binary context or not defined at all. Moreover, the types of biases which are most commonly treated stem from the model building pipeline and can be more broadly categorized as demographic bias, with in- and post-processing bias mitigation strategies being used the most. Revealing these tendencies of definition, instrumentalization and quantification of social and socially constructed concepts helps to contextualize how knowledge and meaning is produced in the field of recommender system research.

Table of Contents

1. Introduction

1.1. Problem statement

The use of recommender systems is vast in our current digital landscape. By limiting options to what is the most relevant according to a carefully chosen algorithm or machine learning model (and an enormous amount of data), they help us choose which music to listen to, which products to buy, which content to consume, which college majors to study, and which people to hire for our company. Recommender systems also shape our culture and have a humongous influence on media literacy. In 2015, Netflix reported that nearly 80% of the streamed hours on their platform came from suggestions made by their recommender system (Gomez-Uribe & Hunt, 2016). However, recent work found that these automated decision-making processes are prone to a set of biases and concerns which are specific to recommender systems (Gössl, 2023). Amazon's hiring tool from which was fed data from the previous decade, including a significantly higher rate of male candidates. The lack of diversity in the training data resulting in the discrimination of female applicants (Gershgorn, 2018). Similarly, the COMPAS program, which is currently still in use by several US-States, was found to deem Black[1] candidates as having a higher re-offense risk than white candidates. The program inherited the judge's bias towards Black candidates, making it less likely for them to be bailed out, as well as making bail more expensive (Flores et al., 2016). The rise of phenomena like these lead to the development of Algorithm Fairness, a new research area dedicated towards developing definitions and measures of fairness, as well as advancing bias mitigation strategies (Ashokan & Haas, 2021). Furthermore, it lead to the founding of conferences such as the ACM Conference on Fairness, Accountability, and Transparency (FAccT).

[1] This thesis engages in the practice of capitalizing the term 'Black' (as well as not capitalizing 'white') when referring to a group's or an individual's racial/ethnic/cultural identity. Note that the concept of 'race' has different meanings for different people, as well as historical implications, and it is not easily translated into other languages. I am writing this thesis from my position as a white European person and aim to use the most inclusive as well as culturally sensitive language to my current knowledge.

Furthermore, I am following Mike Laws reasoning behind this practice, which was described in the *Columbia Journalism Review* style guide as follows: "we capitalize *Black*, and not *white*, when referring to groups in racial, ethnic, or cultural terms. For many people, *Black* reflects a shared sense of identity and community. White carries a different set of meanings; capitalizing the word in this context risks following the lead of white supremacists." Laws, M. (2020). Why we capitalize 'Black' (and not 'white'). *Columbia Journal Review*. https://www.cjr.org/analysis/capital-b-black-styleguide.php

1.2. Research Questions and Related Work

I am currently in the midst of my bachelor program of Artificial Intelligence at the Johannes Kepler University in Linz while also pursuing my master degree in Gender Studies at the University of Vienna. Moreover, coming from a background of Theater-, Film-, and Media Studies, I have found that there exists a very interesting intersection between those three fields of research. The current 'hype' around AI, as well as discourse on the internet, has lead me to pursue this intersection in order to understand AI not just from a technical perspective, but also as a cultural phenomenon embedded within a biased society. My interest in recommender systems as a type of machine learning algorithms is mainly attributed to the subject *Learning from User Generated Data*, which is part of my AI curriculum, where I learned how to build small recommender systems myself. My aim with this thesis is to help advance the ethical implementation and use of machine learning based systems.

As this thesis is an instance of explorative work in the field of Algorithm Fairness, two research questions will guide this analysis in place of a hypothesis:

RQ1: How are the concepts of fairness and gender currently defined in recommender system research?
RQ2: Which biases are most commonly treated and which bias mitigation strategies are currently the most common in the domain of recommender systems?

The first research question was chosen in alignment with my goal in the intersection of Gender Studies and Artificial Intelligence, which is to understand what the current state of research suggests about how to maintain fairness and what it is in the first place. In order to ascertain what fairness is according to researchers in this field, the meanings of terms such as bias and discrimination will be investigated as well, since they are inextricably linked with concepts of fairness. Furthermore, the dimension of gender was chosen as one axis of identity among many due to how prevalent the problem of the gender variable is in computer science. The second research question has been chosen in order to ascertain which strategies are most commonly deployed to combat acts of discrimination, which is most commonly traced back to bias in recommender systems. To conclude, these questions aim to capture the following notion: how researchers define fairness, bias, and gender, as well as how they define action, reflects the priorities and discursive patterns embedded within this scientific community.

This thesis aims to advance knowledge in the field of Algorithm Fairness on a meta-level, investigating and documenting how notions of fairness, discrimination, and gender are represented and operationalized scientific discourse. Pinney et al. have already provided a systematic review of the gender variable in Information Access Systems in *Much Ado About Gender* (Pinney et al., 2023). The results and methods used in their article will be described in more detail in one of the chapters in the Theoretical Background section of this thesis. Furthermore, Scheuerman et al. have provided *Guidelines for Gender Equity and Inclusivity* in the field of Human Computer Interaction (HCI), highlighting how the gender variable is currently (mis)used (Scheuerman et al., 2020). Hamidi et al. have published crucial work on how the construction of gender through Automated Gender Recognition (AGR) can be harmful to vulnerable communities (Hamidi et al., 2018). Devinney et al. have investigated how the gender is constructed in Natural Language Processing (NLP) research (Devinney et al., 2022). My work aims to build upon existing research on the definition of gender in recommender system research, as well as adding the dimensions of fairness, bias, and discrimination as the objects of analysis, in order to sensitize researchers to these notions. It further aims to advance the thoughtful and ethical use of these concepts in computer science.

1.3. Outline

The first main section of this thesis will lay the theoretical foundation to contextualize the later analysis. First, key concepts from the domain of Critical Algorithm and Data Studies will be introduced and exemplified with Safiya Noble's book *Algorithms of Oppression* (Noble, 2018). Next, multiple common fairness measures will be introduced, namely individual fairness, statistical parity, equalized odds and equalized opportunity, disparate mistreatment and disparate impact, calibration fairness, and counterfactual fairness. Additionally, the topic of stakeholders will be discussed. This is followed by a further investigation into notions of fairness with the help of works by Lauren Hoffmann and Os Keyes. In the section on bias definitions, multiple relevant bias frameworks will be introduced, including the one created by Nissenbaum and Friedman in *Bias in Computer Systems* (Friedman & Nissenbaum, 1996). Additionally, ten biases explained in more detail according to their respective pipeline in the recommender system building process: data generation, model building, and user interaction and deployment. This is followed by an explanation of gender concepts in the context of computer systems with the help of works by Judith Butler, Keyes et al., and Pinney et al. The last section of the theory-chapter will be dedicated towards explaining bias mitigation strategies

and their respective pipelines. The methodological chapter of this thesis will be based upon Mayring's *Qualitative Content Analysis*, more specifically, a structuring or deductive content analysis conducted on 24 research papers in the field of recommender system fairness (Mayring, 2014). This chapter includes an explanation on how the research questions are linked to the theoretical foundations of the previous chapter, and it describe the sampling process as well as provide the codebook which was used for the analysis. The following chapter will then present the results of the content analysis with respect to the dimensions of fairness, discrimination, and gender, as well as bias and bias mitigation strategies in recommender system research. These results will then be discussed in detail in the next chapter, which will also provide thoughts on the fairness-accuracy trade-off. Lastly, in the conclusion of this thesis, the research questions will be answered explicitly, as well as potential limits and prospects based on the study conducted within this work. A table containing all categorizations will be provided in the appendix for the sake of scientific transparency.

2. Theoretical Background

2.1. Critical Algorithm and Data Studies

Critical Algorithm and Data Studies is an interdisciplinary field of study which is concerned with examining the cultural and ethical challenges that arise when working with algorithms and data. Researchers in this field investigate issues such as transparency, accountability, and explainability in technological systems. The structure of this section will be mainly based on the reading list published by the *Social Media Collective* on the topic of Critical Algorithm Studies (SMC). This collective is a social science and humanities research network, which is part of the Microsoft Research labs in New York and New England. They provide resources based on empirical and critical methods exploring the cultural and political dynamics embedded in technologies, as well as their consequences.

This thesis aims to analyze research on recommender systems from a perspective of Critical Algorithm and Data Studies. Raghuwanshi defines recommender systems (also synonymously referred to as recommendation systems, platforms, or engines) as "software tools and techniques that provide suggestions for items that are most likely of interest to a particular user." (Ricci et al., 2022, p. 1) The term 'item' denotes the object that is being recommended, which can be anything from movies to news articles to other users. Ricci et al. further describe that "RSs are primarily directed at individuals who lack sufficient personal experience or competence to evaluate the potentially overwhelming number of items that a website may offer" (ibid.) and that recommendations are often personalized to the point of being completely different experiences for different users. Non-personalized recommendations are, they argue, simpler to generate and typically include top-10-examples of items such as movies or books. While being a useful tool in cases where not much information is available, the authors state that non-personalized recommender systems "have not been the primary focus of the RS research." (ibid., p. 2)

Personalization, social sorting, and discrimination via computer systems in general has already been discussed several decades ago by many theorists and researchers, such as Batya Friedman and Helen Nissenbaum, authors of *Bias in Computer Systems* (Friedman & Nissenbaum, 1996). Furthermore, one core concept in Critical Algorithm and Data studies expresses the idea that algorithms are not objective entities and that they are inherently human in their design, as Nissenbaum explained in *How Computer Systems Embody Values* (Nissenbaum, 2001).

Another position present in Critical Algorithm and Data Studies is the critical theory approach which contextualizes algorithms and data collection within capitalism, bringing the aspect of surveillance to the forefront. Newer viewpoints of this include the examination of the concept of the 'Web 2.0', which describes the phenomenon of increasing collaboration and content-creation on user-side. David Beer contested the idea that this increasing mode of collaboration is a sign of 'empowerment' and 'democratization' in *Power through the algorithm? Participatory web cultures and the technological unconscious* (Beer, 2009). The next section is dedicated towards exemplifying some of these key concepts with Safiya Umoja Noble's book, *Algorithms of Oppression. How Search Engines Reinforce Racism.*

2.1.1. Algorithms of Oppression

In 2018, Noble published her book and coined the term 'technological redlining', which aims to describe the ways in which algorithmic decisions can "reinforce oppressive social relationships and enact new modes of racial profiling" (Noble, 2018, p. 1). Outraged by the top-ranked results of a Google search for 'black girls' which immediately produced link to porn websites, Noble investigates how a biased set of search algorithms privilege whiteness and actively discriminate against Women of Color. The author uses textual and media analyses on search engines as well as research on paid online advertising to expose the racist structure of discoverability on the Internet. The choice of the term 'redlining' originates on the fact that Black or Latinx individuals are more likely to pay higher interest rates or premiums if they live in low-income neighborhoods. Similarly, Noble argues that discrimination based on race is inscribed in computer code and decision-making tools such as search engines and recommender systems. On this note, Noble believes that "artificial intelligence will become a major human rights issue in the twenty-first century" (ibid.) and that neither data nor algorithms are "benign, neutral or objective" (ibid.). In order to understand the issue of biased algorithms from a cultural perspective, Noble highlights the racist tendencies of decision makers in Silicon Valley, noting that in 2017, there was an 'antidiversity' manifesto published by former Google engineer James Damore titled *Google's Ideological Echo Chamber*, which argued that gendered income disparities in tech workplaces can be partly explained by biological differences (such as a higher tendency towards 'neuroticism' in women), that these differences are to be expected, and that the result is not oppressive in nature (Damore, 2017). Although Damore was fired by Google subsequently to the publication of the memo, many Google employees have given Damore thanks privately after the fact (Friedersdorf, 2017). Noble

suggests that events like these contest the paradigm of objectivity in technology, which has been previously defined as a key concept of Critical Algorithm Studies. She highlights that

> "some of the very people who are developing search algorithms and architecture are willing to promote sexist and racist attitudes openly at work and beyond, while we are supposed to believe that these same employees are developing ‚neutral' or ‚objective' decision-making tools." (Noble, 2018, p. 2)

Furthermore, Safiya Noble coins the term Black feminist technology studies (BFTS) and uses it as her approach to Internet research, theorizing it "as an epistemological approach to researching gendered and racialized identities in digital and analog media studies", (Noble, 2018, p. 171) offering "a new lens for exploring power as mediated by intersectional identities" (ibid.). She further explains that

> „BFTS is a way to bring more learning beyond the traditional discourse about technology consumption—and lack thereof—among Black people. Future research using this framework can surface counternarratives about Black people and technology and can include how African American popular cultural practices are influencing non–African American youth.[…] Discourses about African Americans and women as technologically illiterate are nothing new, but dispelling the myth of Blacks / African Americans as marginal to the broadest base of digital technology users can help us define new ways of thinking about motivations in the next wave of technology innovation, design, and, quite possibly, resistance." (Noble, 2018)

In the context of this thesis, this approach is relevant because it exemplifies the phenomenon of the paradigm of objectivity in technology and online platforms, while using an intersectional lens to point out the gaps in that paradigm. While they have different purposes, search engines and recommender systems have many traits in common. They are both examples of Information Access Systems (IAS), (Pinney et al., 2023) and both can provide personalized matches. For example, search results on platforms such as Google are personalized to a degree of high variance between different users' search results, despite a widespread cultural expectation of objectivity of search results (Statt, 2018). Noble's approach highlights how seemingly 'objective' and ubiquitous algorithms, such as those employed by search algorithms, result in an ideologically managed retrieval of content.

Terms such as 'fairness', 'bias' as well as 'gender' are variously defined throughout bias mitigation literature. The following section aims to define these terms in order to analyze the corpus in a comparative manner and outline possible problems that might occur in common definitions (or lack of definition) of these terms.

2.2. Fairness Definitions

Several fairness measures in the context of recommender systems will be presented in this chapter. Although initially, many more fairness measures were researched, the final selection

of fairness measures is based on the most common metrics that appeared in the text corpus of the content analysis. Furthermore, this section will not go into detail of how these fairness measures are defined mathematically, as this master thesis focuses on fairness measures and what issue they intend to solve.

2.2.1. Individual Fairness

What is considered to be 'fair' is, to this day, a widely debated concept, especially in machine learning. Researchers have sought to make definitions of fairness operationalizable in order to be able to evaluate them statistically within computer systems. For example, some definitions of fairness used in machine learning, such as statistical parity, aim to treat a population similarly to one of its protected subgroups. A protected subgroup is a subgroup of a population whose protected or 'sensitive' attributes (gender, race, religion, socioeconomic status, etc.) are protected against discrimination by law. A study by Nyarko et al. has shown that initially, many people at first view the introduction of dimensions such as race or gender into machine learning fairness measures as counterproductive and unethical. There exists a popular belief that disregarding sensitive information about individuals in the process of seems *more fair* and that "any contrary practice is assumed to be morally and politically untenable." (Nyarko et al., 2021, p. 1) However, simply leaving out protected attributes in machine learning in order to achieve 'fairness through unawareness' has been proven to be conceptually similar to the 'color blind' approach in antidiscrimination discourse. This approach allows for other attributes to manifest which correlate with the by-passed protected attributes, ultimately resulting in inadvertent discrimination (Apfelbaum et al., 2010, p. 907). Moreover, Nyarko et al. show that

> "people are generally averse to the use of race and gender in algorithmic determinations of ‚pretrial risk'—the risk that criminal defendants pose to the public if released while awaiting trial. We find, however, that this preference for blinding shifts in response to a relatively mild intervention. In particular, we show that support for the use of race and gender in algorithmic decision-making increases substantially after respondents read a short passage about the possibility that blinding could lead to higher detention rates for Black and female defendants, respectively" (Nyarko et al., 2021, p. 1).

In some cases, the strict notion that sensitive attributes should not be part of the decision-making process leads to a phenomenon Castelnovo et al. called *suppression,* which is the practice of not only bypassing the explicit sensitive attribute, but all information correlating to it. as well. The authors explain that "this approach has the main drawback in the potentially huge loss of legitimate information that may reside in features correlated with the sensitive attribute." (Castelnovo et al., 2022, p. 5)

Although often times conflated, ‘fairness through blindness’ or ‘fairness through unawareness’ are slightly different concepts from individual fairness. Castelnovo et al. describes the nuances of different fairness metrics in *A clarification of the nuances in the fairness metrics landscape* (Castelnovo et al., 2022). In this article, the authors state that individual fairness aims to treat similar individuals similarly and that this similarity between individuals can be calculated with task-specific similarity metrics. They further explain that domain experts are generally needed to figure out which feature is the most useful to conduct the comparison on:

> “the simple idea of using standard similarities, e.g. related to the euclidean distance on feature space, does not take into account the trivial fact that some feature [sic!] are more important than other in determining the relationship of an individual to specific target. Namely, for two applicants for a loan, the difference in income is much more important than the difference in, say, age, or even profession. Thus, judging what does it mean to be similar *with respect to a specific task* is not that simple“ (Castelnovo et al., 2022, p. 5)

In conclusion, individual fairness can be considered as a superset of both ‘fairness through unawareness/blindness’ and ‘fairness through awareness’. Castelnovo et al. clarify that the former latter usually involves the use of target and problem specific similarity metrics, whereas the latter „is a simple recipe that does not depend on the actual scenario.” (ibid.)

2.2.2. Statistical Parity

The subchapters on statistical parity, equalized odds, and equalized opportunity are partially based on the MIT online course on *Exploring Fairness in Machine Learning for International Development* (MIT, 2020). Their content is based on works by (Hardt et al., 2016; Kilbertus et al., 2017), (Wadsworth et al., 2018), (Pleiss et al., 2017), as well as (Verma & Rubin, 2018).

Statistical parity, otherwise known as demographic parity, states that the outcome (e. g. the probability to be hired) is independent of the protected attribute (e. g. gender). Hoffmann describes this process as potentially intervening “in positive (i.e., affirmative) ways“ (Hoffmann, 2019, p. 903) This concept of measurement, which is not specific to machine learning and rather statistics in general, has been proven to be problematic in multiple ways. First, statistical parity cannot account for members with multiple protected attributes (e. g. race and gender) because equal probabilities across all attributes may not be imposable. Ensuring fair treatment of one attribute might mean violating fairness in regards to another attribute or multiple attributes (i. e. subgroup). Secondly, individual fairness may be at stake in order to achieve group fairness. To achieve independence of the attribute, for example by making sure

that members of all gender subgroups are hired at equal probability, the algorithm might choose to drop qualified members in order to achieve statistical parity for unqualified members. This is due to the fact that the 'sweet spot' of low false positives (unqualified members who are hired) and low false negatives (qualified members who are not hired) is difficult to achieve when statistical parity is the only fairness criterion to be considered. And third, Castelnovo et al. mentions that statistical parity might not be the best fairness measure to use in contexts where the differences in behavior between groups of different attributes are likely due to systematic discrimination (Castelnovo et al., 2022, p. 7). They exemplify this with a theoretical dataset which shows that women are more likely to pay back their bank loans than men. In this case, there are two ways in which demographic parity can be used to ensure fairness. One idea would be to remove all information which correlates to the gender variable, generally referred to as "gender information" (Castelnovo et al., 2022, p. 7). This, however, would be another instance of suppression, which has been explained in the previous chapter on individual fairness. Another way would be to enforce demographic parity irrespective of all other information. Castelnovo et al. argue that "if it is true that women repay their loans with higher probability, is it really fair to have demographic parity between men and women?" (Castelnovo et al., 2022, p. 7) and that the bank as a stakeholder has no incentive to grant loans equally to groups with different probabilities of paying them back. Without going into detail about the economic nuances embedded within Castelnovo et al.'s theoretical example, the main question that needs to be highlighted is the question of *why* probabilities like these exist in the first place. This aspect of tracing bias in datasets back to their most probable source needs to be considered first and foremost when aiming for fair prediction.

2.2.3. Equalized Odds and Equalized Opportunity

The criterion of equalized odds is considered to be stricter than statistical parity. Instead of trying to achieve the same average probability among all subgroups with protected attributes, equalized odds aims to only enforce equality among individuals who achieve similar outcomes (MIT, 2020). For example, the probability for a qualified member to be correctly hired (true positive rate) and an unqualified member to be incorrectly hired (false positive rate) should be the same amongst all gender subgroups. A less strict version of this is called equalized opportunity, where only qualified members of each subgroup are treated with the same probability to be hired, i. e. only those who are deemed 'worthy of acceptance' are subjected to the fairness criterion. This means that unqualified members of a subgroup (e. g. men) would not necessarily have the same probability of being denied a job than unqualified members of a

subgroup with a protected attribute (e. g. women). Rejecting members across protected groups 'fairly' is therefore not an issue equalized opportunity is concerned with. This strategy is generally deemed as less interventionalist and more achievable than equalized odds.

2.2.4. Disparate Mistreatment and Disparate Impact

Disparate Impact is a machine learning fairness measure which is rooted in U.S. law. This concept, which was originally coined in the context of legal doctrine with the purpose of encoding 'unintentional bias', was eventually translated into computing. In 1971, the case Griggs v. Duke Power Co. yielded the question whether hiring decisions could be discriminatory without intentionally or explicitly making race a deciding factor (Griggs v. Duke Power Co., 1971). Prior to this case, disparate mistreatment was one of the main legal doctrines to determine discrimination based on a protected attributes. Despite not qualifying for direct discrimination, the business was eventually "forced to stop using intelligence test scores and high school diplomas, qualifications largely correlated with race, to make hiring decisions" (Feldman et al., 2015, p. 259). Since then, disparate impact has become the predominant method to determine indirect and unintentional discrimination, although there is no fixed mathematical formula for defining disparate impact legally (Watson v. Fort Worth Bank & Trust, 1988). When it comes to evaluating fairness in recommender systems and machine learning in general, formulas are necessary, and Feldman et al. introduced an '80-percent-rule', a generalization which was advocated for by the Equal Employment Opportunity Commission (EEOC) in 1978 (CFR, 1978). The authors furthermore note that "disparate impact itself is not illegal; in hiring decisions, business necessity arguments can be made to excuse disparate impact" (Feldman et al., 2015, p. 259). Zafar et al. further differentiate between disparate mistreatment and disparate treatment, the latter corresponding to the "very intuitive notion of fairness: two otherwise similar persons should not be treated differently solely because of a difference in gender." (Zafar et al., 2017, p. 1172) Disparate mistreatment therefore might arise when a classifier yields *different false positive rates or false negative rates* between different gender groups, whereas disparate treatment corresponds to different *decisions* (e.g., hiring decisions) a classifier might make for groups of different genders, even though all of their other (non-sensitive) attributes are the same. Lastly, disparate impact focuses on the output of a decision making process and its *consequences*, as was exemplified with the case of Griggs v. Duke Power Co.

2.2.5. Calibration fairness

Another type of fairness measure deals with calibrating recommendations to individual user's preferences, such as movie genres, other types of item properties, or even other users, depending on the stakeholders in the recommender system. Calibration fairness is a concept which states that a user's recommendations should be related to their interests, and their interests are generally calculated based on their ratings. However, recommendation algorithms are often prone to popularity bias, which is a type of bias which will be explained in the section on Popularity bias. This means that there exists a tendency of recommender systems to recommend popular items "while the majority of other items do not get the deserved attention" (Abdollahpouri et al., 2019, p. 1) Abdollahpouri et al. describe the ethical implications of such a tendency in *The Unfairness of Popularity Bias in Recommendation* as follows:

> "long-tail recommendation can also be understood as a social good. A market that suffers from popularity bias will lack opportunities to discover more obscure products and will be, by definition, dominated by a few large brands or well-known artists […] Such a market will be more homogeneous and offer fewer opportunities for innovation and creativity." (ibid.)

In summary, calibration fairness aims to minimize the disparity between how different users are impacted by popularity bias.

2.2.6. Counterfactual Fairness

Finally, counterfactual fairness is another measure which is considered in cases where a recommender system discriminates based on a protected attribute. In order for a recommendation to be fair, it must be independent from this attribute. To achieve this fairness, a scenario is simulated in which all attributes remain the same but the protected attribute(s) is/are changed. Kusner et al. describe the notion behind this concept as follows: "Our definition of counterfactual fairness captures the intuition that a decision is fair towards an individual if it the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group." (Kusner et al., 2017, p. 1) The authors furthermore recognize that modelling fairness this way does not fully capture the complexity of the problem of discrimination, stating that "fairness should be regulated by explicitly modeling the causal structure of the world. Criteria based purely on probabilistic independence cannot satisfy this and are unable to address how unfairness is occurring in the task at hand." (Kusner et al., 2017, p. 9)

2.2.7. Stakeholders

When it comes to recommender system fairness, there are multiple stakeholders which can be considered, such as users (the entity being recommended to), providers (the 'producer' of the

object of recommendation, or in some cases, the recommended entity itself), and items (the product being recommended) involved in recommendation. There are cases in which recommendation must be done bilaterally, such as employment recommendation. In these cases, since all parties must accept the transaction, recommender systems aim to make the most useful recommendation for either side (Burke et al., 2018). Considering multiple stakeholders (as opposed to favoring only users) is therefore one aim of fair recommendation. However, this is not always a simple task. Boratto et al. discussed the problem of the definition of the 'provider' in recommender systems:

> "First, in many cases, there is no direct one-to-one mapping between an item and the individual who has created or offered it (i.e., the provider). Realistic scenarios need to consider items created by more than one provider cooperatively (e.g., a course with two instructors) and how the sensitive attributes are associated to the involved providers. It can be even difficult to come up with a one-to-many mapping for items offered by an entity not directly linked to individuals (e.g., a training com- pany providing an online course)" (Boratto et al., 2021, p. 426).

The second problem further investigates the difficulty of mapping protected attributes to multiple, potentially undefinable providers:

> „Second, the fact that an item might have more than one provider behind it poses the problem of how to model the representation of a providers' sensitive attribute, when considering that item (e.g., how each gender is represented in a given item), based on the individuals associated to it. Linking a unique variable, either binary or multi-class, discrete or continuous, to a sensitive attribute of a provider and claim fairness on such a variable is often impractical. More sophisticated solutions should be considered" (ibid.).

Gómez et al. suggest that providers are historically underrepresented as valuable stakeholders in the task of fair recommendation. They make the following argument in the case of teachers on Massive Online Open Courses (MOOC) platforms:

> „Indeed, when their courses are recommended by an algorithm, they receive a certain exposure in the final ranking. Under- or over-exposing, certain providers might generate or exacerbate disparities and affect the opportunities that are given to teachers to offer their services. When these disparities are associated with sensitive attributes, a recommender system unfairly discriminates teachers (*provider unfairness*)" (Gómez, Shui Zhang, et al., 2022, p. 435)

2.3. Investigating Notions of Fairness

These statistical measures have long been examined under their logical as well as practical flaws. Especially in recent years, with more and more cases of discrimination caused by 'flawed algorithms' popping up in media, there have been various researchers, such as Anna Lauren Hoffmann and Os Keyes, who spoke out about the danger of perpetuating inherently violent systems by merely treating its symptoms through fairness measures and bias mitigation in machine learning.

2.3.1. Where Fairness Fails

In her paper *Where Fairness Fails,* Anna Lauren Hoffmann contests three common principles in the antidiscrimination discourse surrounding fairness in computer systems (Hoffmann, 2019). She highlights three main problems regarding mainstream antidiscrimination discourse. First, Hoffmann points out that the discourse focuses mainly on discreet 'bad actors' who are 'at fault' instead of putting the emphasis on a broad systemic hierarchy. Fairness and antidiscrimination are therefore talked about in narrow cause-and-effect terms. Citing Alan David Freeman, Hoffmann points out two problems with this model: First, this model strips discrimination from its broader cultural context, positing discrimination „not as a social phenomenon, but merely as the misguided conduct of particular actors" (Freeman, 1978, p. 1054). Discrimination is therefore rendered detached from society as a whole and moreso „considered an individual and personal trait" (Gotanda, 1991, p. 44). Secondly, the model aims to neutralize „inappropriate conduct on the part of individual perpetrators" (Freeman, 1978, p. 1053) instead of eliminating all contributing factors which have lead to this act of discrimination.

Hoffmann furthermore relates this to discrimination caused by algorithms such as recommender systems: Instead of combatting the issue on a systemic level and admitting to „the structuring role of technology," (Hoffmann, 2019, p. 8f.) individual human designers are deemed to be at fault for a system's shortcomings. However, even taking away the intentionality of individual people often merely shifts the blame from 'bad actors' to 'bad algorithms' or 'bad data':

> "It still permits ignorance of the ways humans and technology co-conspire to not just passively reproduce but actively uphold and reproduce discriminatory social structures, especially in the case of negative externalities 'learned' by, for example, machine learning systems based on subsequent user interaction […]." (Hoffmann, 2019, p. 904f.)

Due to the 'structural turn' in antidiscrimination scholarship (Bagenstos, 2006), there has been improvement in terms of moving away from this 'bad actor' model. However, Hoffmann states that the 'unconscious bias' has been deemed in technology to be completely untangible,

> „as opposed to something that is variously, but systematically cultivated and maintained. The idea that our biases are somehow apart from us yet can infect our decision-making converts them into something akin to what Freeman (1978) sarcastically called our 'ancestral demons' – that is, a possession or invasion for which we are not at fault but which we should nonetheless seek to purge. In this space, unconscious bias training programs pick up where technical fixes leave off: rather than take responsibility for the ways we are daily and actively complicit in reifying culturally-situated violences, we externalize bias and – after a few all-day seminars – count our demons exorcized." (Hoffmann, 2019, p. 905)

The second problem Hoffmann highlights in antidiscrimination discourse is its tendency towards single-axis thinking which centers disadvantage. The first problem which Hoffmann describes in this regard is that discrimination is not mitigated in an intersectional manner (Crenshaw, 1989). Crenshaw was one of the first who described the concept of oppression on multiple axes as something that does not merely stack up but is rather an interlocking of discriminatory processes. A single-axis thinking therefore "explicitly produces vulnerabilities for those who, like Black women, are multiply-oppressed." (Hoffmann, 2019, p. 906) Furthermore, a single-axis thinking which explicitly centers disadvantage fails to address, for example, whiteness and maleness and norms in general treats those concepts as given instead of something that is socially constructed. Hoffmann further elaborates the problem with this tendency:

> „Instead of treating as morally abhorrent those structural processes that unjustly advantage certain groups, the focus on disadvantage forces us into a kind of benevolent—or, worse, patronizing—stance that flattens our understanding of those already relegated to the ‚basement' of the social hierarchy" (Hoffmann, 2019, p. 906).

To this day, data science relating to machine learning fails to address social hierarches in a way that is both intersectional as well as aware of the tendency towards deeming only disadvantaged individuals as 'marked' due to their race or gender. Hoffmann mentions an article by Michael Kearns et al. regarding fairness 'gerrymandering' (Kearns et al., 2017) which highlights the tendency of machine learning to exclusively focus on protected attributes. Kearns et al. oppose this with a set of methods which identify multiple combinations of protected attributes. Hoffmann counters this approach as follows:

> "But this move, while an improvement, still falls short as an ‚intersectional' approach. Intersectionality is not a matter of randomly combining infinite variables to see what ‚disadvantages' fall out; rather, it is about mapping the production and contingency of social categories." (Hoffmann, 2019, p. 906)

Hoffmann further states that even works by researchers which successfully incorporate "the social contingency of difference with which intersectionality is concerned" (Hoffmann, 2019, p. 906), fail to truly address the social justice issues regarding the institutions which employ facial recognition in the first place, as well as the social hierarchy in which these technologies are embedded. She adds that increasing demographic representation in datasets does not address these issues either. Furthermore, approaches in data science which aim to address the tendency of focusing on disadvantage alone has often times shown to be "painfully neutral" (Hoffmann, 2019, p. 907) i.e. by using terms such as 'non-discrimination' and defining it simplistically through "rough parity in false negative and false positive rates across protected groups." (Hoffmann, 2019, p. 907) Hoffmann states that approaches like these unfortunately

ignore the real-world consequences of misclassification, since some groups may have more resources to contest unfair decisions than others. In summary, Hoffmann states that instead of "grappling with the processes that generate patterns of advantage and disadvantage within and across groups, both disadvantage-focused and 'non-discrimination' approaches limit us to solutions that are, at best, reactive and superficial." (Hoffmann, 2019, p. 907)

Lastly, Hoffmann contests the tendency of antidiscrimination discourse to focus on a limited set of goods, rights, opportunities, and resources. This framework relates to technology in that discrimination is thought of in terms of content moderation, hiring and surveillance, consumer protection as well as finance and market manipulation. However, Hoffmann states that "[s]ocial attitudes, for example, play a significant role in shaping persons' well-being in ways that are relevant to the realization of justice, but addressing them is not wholly reducible to matters of redistribution." (Hoffmann, 2019, p. 908) She adds that issues of distribution are also more simple to solve than issues that relate to social attitudes: „Money lost can be replaced and rights violated can be restored, but corporate apologies, subtle tweaks to a system, or even financial compensation ring hollow in the face of attacks on one's dignity." (Hoffmann, 2019, p. 908) Finally, Hoffmann concludes that „data and algorithms do not merely shape distributive outcomes, but they are also intimately bound up in the production of particular kinds of meaning, reinforcing certain discursive frames over others […]." (Hoffmann, 2019, p. 909)

In her conclusion, Hoffmann suggests three steps to tackle these issues (Hoffmann, 2019, p. 910f.). First, in order to overcome these problems on a systematic level, increased attention is needed in the area of algorithmically mediated systems and how they create social orders through their logics of optimization and reduction. Liberal antidiscrimination discourse, argues Hoffmann, tends towards providing 'quick fixes' as well as not understanding how the logics of disadvantage and advantage work. Secondly, a broader understanding is needed when it comes to how our current social hierarchy and the technologies embedded within produce advantage in a systematic and normative way. Only addressing relative disadvantage fails to address the full context which makes discrimination is possible. And last, Hoffmann states that there needs to be more attention towards the fact that algorithms actively mediate and normalize discourses as well as social conditions. After all, these conditions are what makes the specific distribution of goods and resources possible in the first place. By delving into this topic of antidiscrimination discourse, Hoffmann has shown

> "how an uncritical mirroring of the limits of liberal antidiscrimination discourses risks undermining efforts to move beyond talk of 'bad data' and 'bad algorithms' and towards an intersectional commitment to upending the processes by which institutions, norms, systems generate unjust social hierarchies." (Hoffmann, 2019, p. 911)

2.3.2. Terms of Inclusion

In *Terms of inclusion. Data, discourse, violence*,, Hoffmann also contests the idea that discrimination can be simply mitigated through inclusive methods (Hoffmann, 2021). Hoffmann investigates the discourse surrounding inclusive solutions as a method to combat discrimination and argues that it is not only ineffective but also harbors harmful potential due to its tendency towards evading accountability. Posing as a 'quick fix', inclusive solutions generally don't involve empowering individuals to make their own decisions with respect to their safety and visibility. Hoffmann exemplifies this with Tinder's approach to solve its trans-exclusive algorithm by simply adding an option to identify as trans publicly on the platform with an update of the app in 2016. This step towards inclusivity, based on the very system which includes trans people in the first place, simultaneously marks the discrimination problem as solved while also excluding an entire community of trans individuals who don't wish to be visibly marked as trans, who are concerned about their safety on the platform, or whose identity cannot be easily categorized, etc. In summary, Hoffmann states that inclusive methods are often used as an effective emotional appeal but fundamentally change nothing about the system which excludes and discriminates against specific groups of people: "Though purporting to address harms inflicted by data science and technology, inclusion perversely reifies the power advantaged or dominant groups have to recognize and 'bestow humanity' upon the subjugated." (Hoffmann, 2021, p. 3551)

Hoffmann's text furthermore mentions different concepts of violence such as administrative violence, discursive violence, and data violence, as well as coining the term 'datafication'. In the next few paragraphs I will shortly summarize those terms in order to be able to identify them later in the corpus analysis.

2.3.3. Data Violence, Discursive Violence, and Datafication

Administrative violence, originally coined by Dean Spade, aims to describe how administrative systems such as laws and policies facilitate state violence by defining individuals in terms of categories (Spade, 2015, pp. 20-21). Hoffmann extends the term of administrative violence to focus primarily on "processes of classifying, sorting, bounding, labeling, and optimizing enabled by data technologies, both state-run and privately controlled," (Hoffmann, 2021, p.

3541) arguing that this extension is necessary due to the indistinguishability between the private sector and state data collection. Hoffmann furthermore describes data violence as an instance of 'informational power' which was coined by Sandra Braman, (Braman, 2009) and it is defined as "power that underwrites and manipulates the informational bases of other forms of power, such as instrumental, and symbolic power" (Hoffmann, 2021, p. 3541) Instances of informational power have been well documented throughout history. Hoffmann recounts several historical examples of informational power through the instrumentalization of population data, for example, the expatriation of Native Americans as well as the delocation of Japanese-Americans into internment camps. As technology advances, more and more systems have been developed which are based on surveillance and sorting. Simone Browne has exemplified in *Dark Matters. On the Surveillance of Blackness* how data-based digital surveillance systems are built on White Supremacy and actively harm Black communities (Browne, 2015). Hoffmann concludes that, in a landscape of predominantly algorithmically processed data, data violence becomes an even more complicated issue.

Discursive violence is another term coined by Hoffmann, which mediates the concept that violence can be discharged materially as well as discursively by perpetuating specific norms and ways of being, creating the condition which makes othering possible. Datafication describes a process similar to the ways in which militarization operates, which is via material as well as discursive processes. Hoffmann refers to the works of Cynthia Enloe on militarization and gender, which describes how militarization happens not merely on a level with configures bodies materially, but also through discursive transformation (Enloe, 2000). Hoffmann states that it is exactly this discursive process which seems harmless in comparison, however, this is a fallacy: "This discursive transformation is hardly innocuous; rather, it undermines our ability to neatly distinguish between violent and non-violent conditions, as the latter is often contingent on the former in insidious ways" (Hoffmann, 2021, p. 3543).

2.3.4. Counting the Countless

Another problem which needs to be addressed is the logics upon which data science is built. Os Keyes currently researches at the University of Washington's Department of Human Centred Design & Engineering and published an article called *Counting the Countless. Why data science is a profound threat for queer people* in Real Life Magazine in 2019. In this article, Keyes addresses the question whether achieving 'fairness' in data science is even possible.

First, they suggest a definition of the term ‘data science’: “*The quantitative analysis of large amounts of data for the purpose of decision-making.”* (Keyes, 2019) Keyes suggests that data science, however, is inherently based on reduction and generalization, is fundamentally incompatible with the lived reality of trans lives as well as marginalized individuals in general. Since trans existences are built around “fluidity, contextuality, and autonomy”, (Keyes, 2019) attempting to capture data around them is always going to run into two problems: quantitative analysis and the vastness of data. Quantitative analysis is inherently about standardizing and normalizing information. If a survey involves a free-text box as an attempt to be inclusive and counter a binary concept of gender, the person evaluating the data is going to have to make decisions regarding the categorization of gender identities. Keyes exemplifies these series of decisions and their effect as follows:

> “You’re going to have to determine that ‚bigender‘ and ‚nonbinary‘ should be in the same bucket (or shouldn’t). You’re going to have to work out whether you treat trans women and trans men separately, whether you clump them with ‚women‘ and ‚men‘ respectively, what values mean ‚trans woman‘ in the first place — you are going to lose definition. You are going to have to make judgment calls on where the similarities are and what is (and is not) equivalent. You have to, because quantitative methods work on the assumption that you’ll have neatly distinguishable buckets of values.“ (Keyes, 2019)

Quantitative analysis therefore inherently diminishes queer identities by forcing categories onto them, rendering ‘inclusive approaches’ not just useless, but rather dangerously misleading as well as destructive in terms of keeping definition intact.

Furthermore, data science systems rely on the vastness of data across time, space, and subjects. Their aim is to capture as much of the world as possible for the purpose of decision making. Keyes argues that this process is fundamentally inhumane because it is built to remove inconsistencies and leaves no space for context-specific or non-fixed data by omitting, forcefully categorizing, or ‘fixing’ it. This is where Keyes modifies their opening data science definition into one that is aware of these effects: “*The inhumane reduction of humanity down to what can be counted.“* (Keyes, 2019) Keyes connects this process of inhumane reduction with Anna Lauren Hoffmann’s term ‘data violence’ which has been described in the previous chapter on Data violence, Discursive violence and Datafication.

Lastly, Keyes points out that reform-based solutions to navigate exclusion and reduction through data science have shown to be more harmful than subversive. Keyes uses the example of facial recognition, which has proven to be biased against dark-skinned women in particular (Buolamwini & Gebru, 2018). Reform via inclusion, Keyes argues, is thereby not a valid

approach to combat data violence in a system that is inherently controlling and dangerous to marginalized individuals. However, inclusion is generally deemed as the only logical answer to discriminatory systems. In summary: "data science is fundamentally premised on taking a reductive view of humanity and using that view to control and standardize the paths our lives can take, and it responds to critique only by *expanding the degree to which it surveils us*. […] All reform-based approaches do is make violent systems *more efficiently violent*, under the guise of ethics and inclusion." (Keyes, 2019) Keyes therefore calls upon a radical data science which is decidedly "*not* controlling, eliminationist, assimilatory. A data science premised on *enabling autonomous control of data*, on enabling *plural ways of being.*" (Keyes, 2019)

2.4. Frameworks of Bias

Bias and fairness are closely connected terms, as bias describes the the action which is to be considered unfair (Cambridge, 2024). In the context of machine learning algorithms such as recommender systems, bias is often talked about as having one or multiple origins. In *A Survey on Bias and Fairness in Machine Learning*, Mehrabi et al. explore bias and fairness in various real-world applications. The authors differentiate between bias in data and bias in algorithms as follows:

> „In the cases where the underlying training data contains biases, the algorithms trained on them will learn these biases and reflect them into their predictions. As a result, existing biases in data can affect the algorithms using the data, producing biased outcomes. Algorithms can even amplify and perpetuate existing biases in the data. In addition, algorithms themselves can display biased behavior due to certain design choices, even if the data itself is not biased. The outcomes of these biased algorithms can then be fed into real-world systems and affect users' decisions, which will result in more biased data for training future algorithms" (Mehrabi et al., 2021, p. 3).

It is to be noted, however, that bias can often be traced back to multiple origins, and that the origin of bias can be misattributed. For example, Sara Hooker writes in *Moving beyond ‚algorithmic bias is a data problem'* about how datasets are often times deemed to be at fault for bias, as opposed to the machine learning model itself (Hooker, 2021). This will be discussed more in-depth in the section on algorithmic bias.

Batya Friedman and Helen Nissenbaum pioneered the subject of *Bias in Computer Systems*, which was published in 1996. They define bias as follows:

> „we use the term bias to refer to computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others. A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate" (Friedman & Nissenbaum, 1996, p. 332).

As previously described in the paragraph about Anna Hoffmann's definition of fairness, this definition of bias has since been contested due to its centering of disadvantage as well as the focus on the distribution of resources (in this case: opportunities and goods). However, despite this arguably outdated framing, Friedmann and Nissenbaum's article still holds value, amongst other aspects, in terms of the categorization of bias into different types. Friedman and Nissenbaum further distinguish between three types of bias in computer systems:

2.4.1 Preexisting bias

Preexisting bias is defined as having its "roots in social institutions, practices, and attitudes." (Friedman & Nissenbaum, 1996, p. 334) This definition of bias includes every biased notion that was present before the creation of the computer system, and it might emerge through explicit effort or unconsciously. More specifically, Friedman and Nissenbaum claim that there is a difference between individual and societal preexisting bias: individual bias originates from an individuals such as a client who "embeds personal racial biases into the specifications for loan approval software" (Friedman & Nissenbaum, 1996, p. 334) whereas societal preexisting bias originates from society, ergo from industry or institutions such as legal systems. The accompanying example used here is "gender biases present in the larger society that lead to the development of educational software that overall appeals more to boys than girls" (Friedman & Nissenbaum, 1996, p. 334). It is worth mentioning here that the distinguishing of the individual and the societal is a difficult (and perhaps impossible) task, and it runs the risk of perpetuating the notion that there exist individual biases which have no connection to a societal context, as well as the idea that some biases can be individually mitigated because they aren't rooted in societal structures, as Hoffmann has described in *Where Fairness Fails.*

2.4.2. Technical bias

This bias arises from how technical structures are inherently built. The authors distinguish different categories, such as computer tools, algorithms which have been decontextualized, random number generation as well as the formalization of human constructs. Computer tools include circumstances such as the ordering of items on a screen, which might have to be split into multiple pages, and where the initial page is always given more attention than the other pages. The authors exemplify this as follows: "in a database for matching organ donors with potential transplant recipients certain individuals retrieved and displayed on initial screens are favored systematically for a match over individuals displayed on later screens." (Friedman & Nissenbaum, 1996, p. 334) Another type of technical bias is what Friedman and Nissenbaum call decontextualized algorithms. This type of bias originates in the inability of an algorithm to

make fair decisions due to the necessity of having to rank items in a specific order (such as numerically, alphabetically, etc.) Furthermore, the authors argue that random number generations, being pseudorandom, can also lead to bias. However, their example covers only an edge case in which the randomness is more akin to a weighted coin flip: "an imperfection in a random-number generator used to select recipients for a scarce drug leads systematically to favoring individuals toward the end of the database[…]." (Friedman & Nissenbaum, 1996, p. 334) Finally, the authors describe 'Formalization of Human Constructs' as the category of bias in which data is simplified in order to be processed in an easier way, arguing that it happens "when we quantify the qualitative, discretize the continuous, or formalize the nonformal […]." (Friedman & Nissenbaum, 1996, p. 334) This concept is similar to what Hoffmann described as datafication, however, in this case, the discursive aspect is not highlighted as much as the misuse of technical tools such as quantification and discretization for the purpose of turning humane interpretations of information into processable data. Additionally, as previously illustrated, Os Keyes argues that it is this 'Formalization' aspect which is the inhumane reduction humanity, the fundament on which data science is built upon.

2.4.3. Emergent bias

The third and last category of bias described by Friedman and Nissenbaum is emergent bias, which arises in the process of using the computer system, for example, through user interfaces, since these "by design seek to reflect the capacities, character, and habits of prospective users." (Friedman & Nissenbaum, 1996, p. 335) This type of bias has its roots in changing societal values, norms and knowledge. More specifically, 'New Societal Knowledge' is a subcategory of emergent bias, where new information cannot be integrated into an already existing system, rendering it biased. This is distinguished from a 'Mismatch between Users and System Design', where the userbase which the system was intended for differs from the actual userbase using the system. This is further categorized into a mismatch due to expertise, where the userbase is assumed to have a different knowledge base than they actually do, and a mismatch due to different values, where a technical design has embedded values which do not represent the values of the actual userbase.

Although worth mentioning, Friedman and Nissenbaum's framework does not suffice for this thesis' in-depth research paper analysis because of two main reasons: Firstly, the research corpus comprises of papers published in the years 2017-2023. Most of these papers do not frame the biases they aim to mitigate in the terms that Friedman and Nissenbaum introduced

in1996. It might be more accurate to think of current bias frameworks as a mix or combination of the three main categories which were previously introduced. Some of the biases treated also were not an issue during the time Friedman and Nissenbaum's paper was published and only became possible with modern model building practices. Secondly, Friedman and Nissenbaum's framework does not transfer easily to the domain of recommender systems. I have therefore chosen a bias framework which encompasses three stages of the recommender system building process. It contains biases which emerge during data generation phase, during the model building period, and lastly, during the deployment and user interaction phase. This categorization was based on the pipeline categorization done by Ashokan and Haas and was done to make the categorization of research papers more efficient and domain-specific (Ashokan & Haas, 2021). All of the following biases were featured in their paper, except for filter bubbles, which is an additional bias/phenomenon which was added to the category of user interaction and deployment biases.

2.4.4. Data Generation Bias

i. Historical Bias

Ashokan and Haas define this bias as follows: "Bias in data generation process due to the already existing bias from socio-technical issues." (Ashokan & Haas, 2021, p. 4) This arguably broad definition aims to comprise all biases which stem from systemic societal power structures. It might be worth arguing that technically, there does not exist any bias that is untethered from societal structures, making all bias 'historical'. But in the context of this research analysis, there are papers which aim to mitigate the type of bias which is already present in the structure of the data used for the recommender system. Whether or not a paper aims to mitigate historical bias is also an issue of framing by the authors of a research paper. Generally, research papers claiming to "mitigate fairness concerns that go beyond individual users and items towards more systemic biases in recommendation" (Wu et al., 2022, p. 1) fall into this category.

In 2021, Suresh and Guttag published a paper on the contextualization of harm caused my Machine Learning models, titled *A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle*. The authors of this paper define historical bias in differentiation to biases which occur during sampling, stating that:

> „Historical bias arises even if data is perfectly measured and sampled, if the world *as it is* or *was* leads to a model that produces harmful outcomes. Such a system, even if it reflects the world accurately, can still inflict harm on a population. Considerations of historical bias often involve evaluating the

> representational harm (such as reinforcing a stereotype) to a particular group.“ (Suresh & Guttag, 2021, p. 4)

The term ‘representational harm’ is referring to the type of harm caused “(when certain people or groups are stigmatized or stereotyped“, (Suresh & Guttag, 2021, p. 2) differentiating it from allocative harm, which happens “when opportunities or resources are withheld from certain people or groups“, (ibid.) stating that allocative harm is “typically the type of harm that we think and hear about, typically the type of harm that we think and hear about, because it can be measured and is more commonly recognized as harmful.“ (ibid.) Allocative harm has also been previously mentioned in the context of Hoffmann’s critique of the framing of discrimination in contemporary discourse, where the focus lies on a matter of distribution of resources as opposed to a systematic change of social attitudes. Suresh and Guttag however, explain in the context of their framework that even if harm is not as obvious as with allocative harm, representational harm still exists. Their focus lies on these two main types of harm in order to make the extent of harm caused by historical bias clear.

To understand historical bias fully, it is necessary to address the debate which its definition is embedded in. In the introduction of their paper, the authors reference the debate whether biased machine learning is purely based on biased data. They address a social media post by Turing Award winning computer scientist and machine learning expert Yann André LeCun, who posted on X (formerly known as Twitter) in 2020:

> “ML systems are biased when data is biased. This face upsampling system makes everyone look white because the network was pretrained on FlickFaceHQ, which mainly contains white people pics. Train the *exact* same system on a dataset from Senegal, and everyone will look African.“ (LeCun, 2020)

Suresh and Guttag counter this argument, stating that, while biased data *does* exist, the harm caused by machine learning systems cannot be solely attributed to flawed data distributions and labelling, since

> “[t]his process is long and complex, grounded in historical context and driven by human choices and norms. Understanding the implications of each stage in the data generation process can reveal more direct and meaningful ways to prevent or address harmful downstream consequences that overly broad terms like ‘biased data‘ can mask.” (Suresh & Guttag, 2021, p. 1)

The authors prove this by counterexample. When data is sparse and underrepresents a protected grpup, simply getting ‘more data’ from the same underlying biased distribution will not result in a less biased machine learning model. Sometimes, getting more data is useful: more medical data on female heart attack patients can result in a more consistent prediction model. Acquiring

more data when the data is based on human assessment (labeling) of quality in order to determine whether an applicant for a job is qualified for hiring might not be as useful. In the author's hypothetical example, the model output was still biased against women, since "using human assessment of quality as a label to estimate true qualification allowed the model to discriminate by gender, and collecting more labelled data from the same distribution did not help." (ibid.)

ii. Representation Bias

Suresh and Guttag define representation bas as follows: "Representation bias occurs when the development sample underrepresents some part of the population, and subsequently fails to generalize well for a subset of the use population." (Suresh & Guttag, 2021, p. 4) Representational bias may arise in three distinct ways. Firstly, the target population (users of an app, for example) defined may not match the use population. This can, for example, mean that data describing specific attributes of one city may not be representative for another city. Furthermore, the target population may contain under-represented groups, which may lead to a model being less robust to that specific subgroup. The authors exemplify this with a target user base of ages 18-40, where 5% of those users are pregnant. The model is less likely to learn about these 5%, making the model less robust for the pregnant use population, even if target and use population are exactly the same. And lastly, the sampling method used to sample from the target population could be limited or uneven. For example, this can happen when medical data is only available for a subgroup of the target population, such as patients whose medical condition is deemed as 'serious enough' to require further screening. This may result in a skewed representation of the target population. The authors add that "[i]n statistics, this is typically referred to as sampling bias." (Suresh & Guttag, 2021, p. 5)

Suresh and Guttag exemplify representation bias by explaining that ImageNet, a visual database often used for object recognition software, which have been hand-annotated in order for the model to be able to train and recognize the objects in a given image. The authors state that "ImageNet does not evenly sample from this target population" (ibid.) and that „approximately 45% of the images in ImageNet were taken in the United States, and the majority of the remaining images are from North America or Western Europe. Only 1% and 2.1% of the images come from China and India, respectively." (ibid.) This leads to a skewed representation of the target population, as has been shown by (Shankar et al., 2017) and to a significantly worse performance when it comes to "classifying images containing certain

objects or people (such as ‚bridegroom‘) when the images come from under-sampled countries such as Pakistan or India.“ (Suresh & Guttag, 2021, p. 5)

iii. Simpson's Paradox

Ashokan and Haas define the Simpson’s Paradox as „[b]ias from the difference in behavior of population sub-groups when aggregated and taken individually.“ (Ashokan & Haas, 2021, p. 4) In summary, if a population shows certain statistical trends in terms of features, it doesn’t necessarily mean that its sub-populations exhibit similar trends. In fact, they may show opposite trends than those of the overall population. The same may also be true in the other direction: just because the sub-groups of a population all individually exhibit similar trends, doesn’t mean that the population as a whole exhibits those tendencies.

Xu et al. exemplify the Simpson’s Paradox in the context of recommender systems with a simple toy dataset. In this example, a group of 100 users have been categorized into two subgroups in terms of a protective trait (gender, age, race, etc.) (Xu et al., 2023, p. 236). There are also two items which can be interacted with and potentially rated with ‘like’ or ‘dislike’. 50/80 users from group A liked item 1 and 20/30 liked item 2. From group B, 10/20 users liked item 1 and 39/70 users liked item 2. Overall, item 1 seems to be the favorable item since 60/100 users liked it, as opposed to 58/100 likes for item 2. This means that item 2 will be recommended to each of the user groups, since it seems to be an overall favorite. However, as has been shown, both user groups A and B individually actually prefer item 1 to item 2.

Although the Simpson’s paradox is a matter of statistical bias and can therefore arise in a variety of domains, it is a crucial bias to consider when it comes to recommender systems and the potential harms they exhibit, especially when groups of users are categorized into groups according to a protected trait.

2.4.5. Model Building Bias

iv. Popularity Bias

This bias occurs due to over-exposure of popular items (Ashokan & Haas, 2021, p. 4). This problem is related to the ‘long-tail / short-tail’ problem in machine learning, which has been summarized by Musto et al. in *Fairness and Popularity Bias in Recommender Systems. an Empirical Evaluation* (Musto et al., 2021). The long-tail phenomenon, which is based on Zipf’s law, refers to a specific way in which data is distributed and observed. Zipf’s law is found in

many distributions such as wealth of a population or word frequency in a language. It states that the second most 'popular' (e.g. frequent) item in a data corpus is half as popular as the most popular item. The third most popular item is about a third as popular as the first item, and such, the popularity lessens exponentially for each consecutive item. Consequently, there exists an imaginary threshold which separates the very small amount of relatively popular items (the 'short tail' of the distribution) from the very large amount of relatively unpopular items (the 'long tail' of the distribution). This is relevant for recommender systems, since items in the 'short tail' are more likely to be recommended, resulting in what Fleder and Hosanagar coined as the *Blockbuster effect* (Fleder & Hosanagar, 2009). In offline evaluations of recommender systems, this effect often aids as a baseline for accuracy, but it nonetheless limits the fairness of recommendations, since 'long tail' items are being systematically under-recommended. Musto et al. state that it is important for a good recommender system to achieve balance between items of varying degrees of popularity (Musto et al., 2021, p. 2f.).

v. Algorithmic Bias

Ashokan and Haas simply describe this bias as the type of "[b]ias that gets added purely by the algorithm when the input data is unbiased." (Ashokan & Haas, 2021, p. 4) It is therefore explicitly distinguished from historical bias and other data sampling biases, since this definition implies that there is no bias occurring in the sampling process, resulting in an unbiased distribution of the data.

The task of finding out where the bias in a biased recommender system comes from is itself a difficult problem. Generally, stating that unbiased data exists is part of a debate that is active to this day, as has been described previously in the chapters summarizing Anne Hoffmann's perspective on 'fairness' discourse as well as Os Keyes' definition of data science.

However, the claim that model bias is *purely* a data problem and that algorithms themselves are impartial is just as much as a questionable statement as the previous one. Sara Hooker addressed the question in her article *Moving beyond ‚algorithmic bias is a data problem'*:

> „A surprisingly sticky belief is that a machine learning model merely reflects existing algorithmic bias in the dataset and does not itself contribute to harm. Why, despite clear evidence to the contrary, does the myth of the impartial model still hold allure for so many within our research community? Algorithms are not impartial, and some design choices are better than others. Recognizing how model design impacts harm opens up new mitigation techniques that are less burdensome than comprehensive data collection." (Hooker, 2021, p. 1)

Hooker disagrees with the notion of impartial algorithms, stating that "some design choices are better than others." (ibid.) Furthermore, she says that "[a] model can fulfill an objective in many ways, while still violating the spirit of said objective." (ibid.) She also explains that a more nuanced understanding of the contributing factors to algorithmic bias is crucial since it dictates the place where harm mitigation takes place. She adds that

> „[i]f algorithmic bias is merely a data problem, the often-touted solution is to de-bias the data pipeline. However, data ‚fixes' such as re-sampling or re-weighting the training distribution are costly and hinge on (1) knowing *a priori* what sensitive features are responsible for the undesirable bias and (2) having comprehensive labels for protected attributes and *all* proxy variables." (ibid.)

Hooker states that satisfying both aspects in the real word is, more often than not, infeasible. The aforementioned *a priori* brings about many difficult problems because, among other things, agreeing on a standard taxonomy is a hard problem, since "categories attributed to race or gender are frequently encoded in inconsistent ways across datasets." (ibid.) and therefore, protected attributes may be "perceived as intrusive leading to noisy or incomplete labels." (ibid.)

The problem of long-tail and short-tail distribution is also taken up in Hookers article. Hooker explains the relationship between this distribution phenomenon and algorithmic bias as follows:

> "A key reason why model design choices amplify algorithmic bias is because notions of fairness often coincide with how underrepresented protected features are treated by the model. […] algorithmic bias a model learns can be attributed to the relative over-and-under representation of a protected attribute within a dataset category. Most real-world data naturally have a skewed distribution […], with a small number of well-represented features and a 'long-tail' of features that are relatively underrepresented. The skew in feature frequency leads to disparate error rates on the underrepresented attribute. This prompts fairness concerns when the underrepresented attribute is a protected attribute but more broadly relates to the brittleness of deep neural network performance in data-limited regimes. Understanding which model design choices disproportionately amplify error rates on protected underrepresented features is a crucial first step in helping curb algorithmic harm." (Hooker, 2021, p. 1f.)

In summary, Hooker concludes that the discussion around algorithmic bias and whether or not it is purely a data problem might be related to the broader social phenomenon of diffusion of responsibility. She defines the term as a "socio-psychological phenomenon where an individual abstains from taking action due to the belief that someone else is responsible for intervening." (Hooker, 2021, p. 3) She connects this term with a common practice in computer science, since this type of diffusion defines what is 'out of scope' and what isn't, and that many problems are delegated until labeled as 'somebody else's problem'.

vi. Demographic Bias

According to Ashokan and Haas, demographic bias is defined as "[b]ias that occurs from users of different demographic groups (age and gender) being treated differently." (Ashokan & Haas, 2021, p. 4) The problem with the definition of demographic bias as a separate form of bias is that one could argue that *all* bias is demographic if it affects users belonging to a protected group. For example, Drozdowski et al., have defined general bias in algorithms as follows: "an algorithm is considered to be biased if significant differences in its operation can be observed for different demographic groups of individuals (e.g., females or dark-skinned people), thereby privileging and disadvantaging certain groups of individuals." (Drozdowski et al., 2020, p. 89) which implies that bias in algorithms always comes with an effect on certain demographic groups.

The reason why demographic bias is still considered a separate category from representation bias is because the definitions some papers use to define the type of bias that is treated with their proposed bias mitigation strategy mediate different areas of focus. Ekstrand et al., for example, argue that the attention towards mitigating demographic bias is important in order to cater to different needs a demographic might have. However, as has been argued in the case of representation and popularity bias, the needs of the largest subgroup will usually prevail. Ekstrand et al. explain that,

> "[i]f other subgroups have different needs, their satisfaction will carry less weight in the final analysis. This can lead to a misguided perception of the performance of the system and, more importantly, make it more difficult to identify how to better serve specific demographic groups." (Ekstrand et al., 2018, p. 2)

2.4.6. Deployment and User Interaction Bias

vii. Social Bias

Ashokan and Haas describe this bias as the type of „[b]ias that occurs from other people influencing ones judgment" (Ashokan & Haas, 2021, p. 4). Baeza-Yates exemplifies this via collaborative ratings, where a user comes across an item which they do not like, but which has already been rated by other users with a high score. This increases the user's probability to rate this item with a high score as well, since this new information may lead them into thinking that their judgement is "too harsh." (Baeza-Yates, 2018, p. 60) This bias is also often times referred to as the 'herding effect' or as social conformity. Social bias also occurs when users are influenced by popular users. Olteanu et al. have explored this bias, stating that "[s]ocial platforms are not closed systems. They are open to external influences that affect the makeup

of the user populations enticed to each platform, as well as their interests and activities.“ (Olteanu, p. 15, Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries) Wang & Wang also mention the inherent paradox in social bias as well as the discrepancy between the perceived value of an item based on collective ratings and an item’s intrinsic value:

> “Our society is increasingly relying on digitalized, aggregated opinions of individuals to make decisions (e.g., product recommendation based on collective ratings). One key requirement of harnessing this ‚wisdom of crowd‘ is the independency of individuals' opinions; yet, in real settings, collective opinions are rarely simple aggregations of independent minds.“ (Wang & Wang, 2014, p. 1)

viii. Interaction Bias

This bias is created due to interaction differences between users and a recommender system. Baeza-Yates explains that user interaction is highly idiosyncratic and can differ on many levels, such as eagerness to click or moving where the user looks:

> “Mouse movement is a partial proxy for gaze attention and thus a computatio ally inexpensive replacement for eye tracking. Some of us may not notice the scrolling bar, others prefer to read in detail, and yet others prefer just skim. In addition to the bias introduced by interaction designers, users have their own self-selection bias.” (Baeza-Yates, 2018, p. 60)

This self-selection bias is a cultural and cognitive bias related to how users choose answers aligned with their already existing beliefs when it comes to queries in web search engines. Baeza-Yates furthermore mentions the complexity of interaction bias when it comes to developers isolating them and separating them from different cultural and cognitive biases, stating that “biases cascade and depend on one another, implying that Web developers are always seeing their combined effects.“ (ibid.)

ix. Ranking Bias

Ashokan and Haas describe this bias as “[b]ias that occurs from top-ranked results being more clicked upon.“ (Ashokan & Haas, 2021, p. 4) Mehrabi et al. attribute ranking bias to being a type of user-interaction bias, since the popularity of higher-ranked result is “not because of the nature of the result but due to the biased interaction and placement of results by these algorithms“ (Mehrabi et al., 2021, p. 4). Similarly, Baeza-Yates explains that the top-ranked result in a search engine which lists content in relevant order from top to bottom will always attract more clicks than others, not just because it is more relevant, but also because it is ranked in one of the first positions. The author suggests that “[t]o avoid ranking bias, Web developers need to de-bias click distribution so they can use click data to improve and evaluate ranking algorithms. Otherwise, the popular pages become even more popular.“ (Baeza-Yates, 2018, p.

59) Ranking bias is often used interchangeably with the term ‘position bias’, which is also sometimes described as a type of bias where higher ranked items on a list are favored over lower ranked items (Azzopardi, 2021, p. 31). Baeza-Yates further notes that cultural factors need to be considered as well when it comes to position bias, such as reading from left to right and top to bottom in western cultures (Baeza-Yates, 2018, p. 59). Position bias poses a significant problem for recommender systems such as search engines. Forbes reported that less than 6% of website clicks happen on the second page of a search engine (Shelton, 2017). A 2023 study found that only second-page clicks on Google only happens in 0,63% (Dean, 2023).

iv. Filter Bubbles

This chapter explains the concept of filter bubbles on internet platforms and their connection to recommender systems. The term ‘filter bubble’ was coined by author and activist Eli Pariser in 2011 in *The Filter Bubble. What the Internet is Hiding from You* (Pariser, 2011b). In a Q&A with the author, Pariser argues that the filter bubble is an effect of the attention economics on the Internet, which inevitably leads to intellectual and social fragmentation and isolation:

> “Were [sic!] used to thinking of the Internet like an enormous library, with services like Google providing a universal map. But thats [sic!] no longer really the case. Sites from Google and Facebook to Yahoo News and the New York Times are now increasingly personalized based on your web history, they filter information to show you the stuff they think you want to see. That can be very different from what everyone else sees or from what we need to see. Your filter bubble is this unique, personal universe of information created just for you by this array of personalizing filters. Its invisible and its becoming more and more difficult to escape“ (Pariser, 2011a).

Baeza-Yates discussed the topic of filter bubbles in Bias on the Web as an example of second-order bias, which is the type of bias that is created through a cycle of user interaction and platforms responding to input. The author states that personalization algorithms such as filter bubbles create a bias that does not affect content on the internet itself but rather which type of content the user is exposed to (Baeza-Yates, 2018, p. 56). Baeza-Yates explains how the main issue seems to not lie in personalized algorithms per se, but rather in how user interaction data is used:

> „If a personalization algorithm uses only our interaction data, we see only what we want to see, thus biasing the content to our own selection biases, keeping us in a closed world, closed off to new items we might actually like. This issue must be counteracted through collaborative filtering or task contextualization, as well as through diversity, novelty, serendipity, and even, if requested, giving us the other side. This has a positive effect on online privacy because, by incorporating such techniques, less personal information is required“ (Baeza-Yates, 2018, p. 60f.).

The question of how collaborative filtering affects the filter bubble phenomenon has been explored in a study by Nguyen et al., where the filter bubble has been defined as “a self-

reinforcing pattern of narrowing exposure that reduces user creativity, learning, and connection.“ (Nguyen et al., 2014, p. 677) More specifically, the authors explored whether recommender systems expose users to narrower content over a specific time period, and also how the experience of users who regularly take recommendations differs from those who do not.

The study was conducted with a two-group design on a dataset from a movie recommender system called MovieLens. The results suggested that the group which did not follow recommendations actually had less content diversity over time than the group who did (Nguyen et al., 2014, p. 683). This, at a first glance, suggests that collaborative filtering based recommendation actually lessens the filter bubble effect. However, it is worth mentioning that the study was limited in terms of time since only data form a time span of 21 months was collected. Potential changes in behavior of content recommendation after this time period is not included in the study, and the authors state that “[e]ventually the content diversity of the Following Group may become less than that of the Ignoring Group. However, this is an issue for future work.“ (ibid.) Furthermore, the study concluded that at the end of the observed time periods, “the content diversity of both groups is reduced.“ (ibid.)

Due to the limitations of the study conducted by Nguyen et al., it is not sufficient to make conclusive statements about the effect of *all* recommender systems on filter bubbles. The main observation that needs to be pointed out, however, is that, although “taking recommendations [based on collaborative filtering] lessened the risk of a filter bubble“, (Nguyen et al., 2014, p. 685) content diversity seemed to still decrease for both user groups. This suggests that the ways in which filter bubbles form is more complex than simply ‘following recommendations’, and instead, is a process which comes with consuming content on Internet platforms due to a multitude of cognitive and socio-economic factors. Nguyen et al. conclude:

> “This begs the question - is there a ‚natural‘ narrowing effect over time, at least in the domain of movies? After all, we form habits based on what we’ve watched recently, and as we watch more, we solidify our preferences. In the movie domain, we face the additional possibility that the best movies are relatively diverse in content, but limited in number; once we get through those, we turn to newer movies closer to our comfort zone. If this is true – if there is a natural tendency to narrow our consumption of movies (or other media) over time – then collaborative filtering-based recommenders appear to help mitigate the tendency, and thus may play a broadening role.“ (ibid.)

The naturalization of the filter bubble effect in this quote is not to be disregarded. It is worth highlighting that it is exactly this naturalization of cognitive biases which have been criticized by Os Keyes and Hoffmann in previous chapters. It is not by chance that Pariser explicitly

doubles-down on the *deliberateness* of the filter bubble effect being an effect of an Internet economics based on advertisement:

> "The rush to build the filter bubble is absolutely driven by commercial interests. Its becoming clearer and clearer that if you want to have lots of people use your website, you need to provide them with personally relevant information, and if you want to make the most money on ads, you need to provide them with relevant ads. This has triggered a personal information gold rush, in which the major companies Google, Facebook, Microsoft, Yahoo, and the like are competing to create the most comprehensive portrait of each of us to drive personalized products. Theres [sic!] also a whole behavior market opening up in which every action you take online every mouse click, every form entry can be sold as a commodity." (Pariser, 2011a)

2.5. Gender Definitions and Operationalization

2.5.1. Gender Trouble and Bodies That Matter

Judith Butler, who is currently a researcher at the University of California, Berkeley, has had a massive influence on the research field of critical theory as well as gender and sexuality studies. *Gender Trouble* and *Bodies That Matter* count as some of their most influential work. The following section aims to shortly summarize the key concepts of these books.

Butler argues that gender does not exist as a natural, objective reality: „Gender reality is performative which means, quite simply, that it is real only to the extent that it is performed" (Butler, 1990, p. 278) Gender is therefore not tied to 'material facts' or physical bodies:

> „Because there is neither an 'essence' that gender expresses or externalizes nor an objective ideal to which gender aspires; because gender is not a fact, the various acts of gender creates the idea of gender, and without those acts, there would be no gender at all. Gender is, thus, a construction that regularly conceals its genesis" (Butler, 1990, p. 273)

Gender is thereby retroactively produced by performative acts, which constitutes the agent as the object (as opposed to the subject) of said acts (Butler, 1990, p. 270). Moreover, Butler argues that they understand those constituting acts „not only as constituting the identity of the actor, but as constituting that identity as a compelling illusion, an object of *belief*" (Butler, 1990, p. 271).

Butler also contests the common feminist distinction between the historical category of gender and the biological category of bodily sex. They argue that sex is also affected by the performative acts which constitute gender and social conventions. Sex is thereby not „a bodily given on which the construct of gender is artificially imposed, but... a cultural norm which governs the materialization of bodies" (Butler, 1993, p. 2f.) Sex, according to Butler, „is an ideal construct which is forcibly materialized through time. It is not a simple fact or static

condition of a body, but a process whereby regulatory norms materialize 'sex' and achieve this materialization through a forcible reiteration of those norms“ (Butler, 1993, p. 2).

In conclusion, Butler applies the postmodern idea of reality being constituted through language on gender as well as on sex. They state that „there is no reference to a pure body which is not at the same time a further formation of that body“ (Butler, 1993, p. 10).

2.5.2. You Keep Using That Word

In *You Keep Using That Word: Ways of Thinking about Gender in Computing Research*, Os Keyes et al. present a list of prompts which aim to sensitize researchers to the complexity of gender by integrating the concept of gender multiplicity as a toolkit (Keyes et al., 2021). The authors argue that gender is a cultural concept which consists of multiple dimensions and that it needs to be considered which dimensions are more important than others given the specific task. Common pitfalls for researchers are clumping together different aspects of gender and to disregard power relations altogether. They state that "[a] more nuanced (and we would argue, better) approach is for researchers to decompose their use of 'gender' and seek out ways of measuring the multiple concepts relevant to their research that underlie the term." (Keyes et al., 2021, p. 6) They exemplify this by suggesting to use a questionnaire which aims to not just get information about gender identity, but also: perceived gender, perceived level of conformity, internal model of gender, and measure of gender expression.Through this in-depth questionnaire, different relevant combinations of gender may arise:

> "the experiences of those who do not fit such a frame; whose masculinity does not fit their society; whose femininity fits their society but not their community; who are not congruent with gendered expectations, expectations far more complex than can be summarised in a single, categorical variable" (ibid.)

Capturing these experiences is significant, since it is exactly those individuals who do not fit into the status quo which are "in many respects, those who most torque with and are unequal under gendered power—and so those whose experiences must be made most visible in efforts to identify or *address* that inequality." (ibid.) In this context, Keyes et al. follow Butler's approach that the regulation of gender as a binary and normative structure involves the institution of a heterosexuality which is both compulsory and naturalized and further state that research tends to produce a view of gender as homogenous across time and cultures, and that variations are treated as "second-order phenomena" (Keyes et al., 2021, p. 8), even in cases where cultural differences are being acknowledged. They mention that existing gender studies and sociology research show that, even when surveys intend to be inclusive, it is often fundamentally still aligned with western understandings of gender, which leads to the

alienation of Indigenous respondents (Bauer et al., 2017). Keyes et al. furthermore argue that it is necessary for researchers to acknowledge that concepts of gender change, not just on a large-scale cultural level, but on an individual level as well. This includes not just real-life changes in terms of physical spaces and communities, but also internet platforms (Keyes et al., 2021, p. 8).

The authors also mention a tool at the intersection of Artifical Intelligence and gender: gender prediction. Keyes et al. explain that algorithms which aim to predict the gender of a person are generally based on gender as a binary variable as well, treating gender as a generic, static variable used for measurement, "rather than something constructed in ways that fundamentally implicate power" (Keyes et al., 2021, p. 11). This constrains research as well as substantial harm, since research plays a significant role in the production of meaning and could therefore limits the legitimization of "different ways of being and different possible futures and futures" (ibid.).

To conclude, the authors summarize different research questions with the aim to sensitize researchers to the concept of gender. These questions prompt the researchers to define gender themselves as well as ask their participants how they define gender, define the role which gender plays in the research, who and what is left out, and how they are accountable for their work (Keyes et al., 2021, p. 14). They furthermore call for an inquiry into aspects of gender identity which may have been neglected, such as masculinity and cisgenderism.

To further convey what makes gender inclusivity and multiplicity in research important, the authors appeal to a less simplistic view on ignorance in research. They state that ignorance should be understood as more than merely an absence of knowledge, but rather "a more complex view understands ignorance as constituting not simply 'what we do not know' but as an active, socially shaped and cultivated phenomenon" (Keyes et al., 2021, p. 9). In the context of gender, this means that a simplistic view on gender does not simply fail to generate knowledge about gender roles in a certain environment, but rather that "the burden will disproportionately fall on those populations whose identities and experiences are not easily captured" (ibid.) by those who do not fit said gender roles, or whose experiences are, for example, "not easily captured by a model which assumes masculinity and manhood to be one and the same." (ibid.) Lastly, Keyes et al. highlight the role which the production of knowledge plays, and that there *do* exist purposeful practices of ignorance which seek "to remain outside externally imposed frames of knowledge" (Keyes et al., 2021, p. 10). However, there is a

difference between ignorance as an act of resistance and ignorance which is systematically formed in order to oppress minorities. In this context, they quote Star and Bowker on the topic of silence: “Not all silences are benign, nor are all malevolent. Those implemented in the service of erasure are immoral; those created or held for the purpose of reflection, rest and re-thinking are not” (Star & Bowker, 2007, p. 279).

2.5.3. Much Ado About Gender

Pinney et al. have published their research article *Much Ado About Gender* in 2023, investigating how gender is used as a variable in scientific literature in the field of Information Access System research, which includes search engines and recommender systems among other information retrieval systems. Being topically and methodologically similar to this thesis, the article by Pinney et al. poses a significant instance of related work. The authors conducted an inductive content analysis on a corpus of research literature, investigating when gender is an inappropriate or harmful variable to use in research (and when it is not), how gender should be defined and used in cases where it is appropriate, as well as which methods are useful for obtaining gender data and which one’s aren’t (Pinney et al., 2023, p. 269).

First, they exemplify good and bad uses of the gender variable in Information Access System research. As a negative instance of the gender variable use, Pinney et al. mention the advertisement strategy employed by the companies KFC and Baidu in Beijing, China in 2016 (Etherington, 2016). This strategy involved inferring variables such as gender, age and ‘beauty’ from customers in order to personalize menus for them. Pinney et al. state that it is unclear how these variables are relevant in terms of food choice, however, they state that “what we do know is that the system presents a new avenue for massive collection of facial images and purchasing patterns, which could be used by Baidu to monetize other aspects of social and economic life in China.” (Pinney et al., 2023, p. 270) An example for a positive instance of the gender variable was found in an audit conducted by Spotify in 2020 (Epps-Darling et al., 2020). They investigated how female artists, which are systematically underrepresented in the music industry, were represented and made visible on the platform. Furthermore, they found that the platform’s discovery tools nudge users to listen to more female artists than they would ‘organically’ (without recommendations). In this instance, the research goal was to audit a system for the purposes of fairness, which was found to be more a compatible goal when it comes to an ethical use of the gender variable. The gender variable (as well as other demographic data) is, however, used in IAS for a variety of purposes. They explain that data

brokers often times collect and sell gender data to companies. There exists gender data embedded within movies, which is commonly represented using a gender affinity axes, which expresses whether a movie is 'geared towards' men or women. Furthermore, recent work aims to understand how IAS may treat some users unfairly based on their gender, differentiating between whether they are treated differently "as producers of the information being retrieved […], or as the subjects of that information […]." (Pinney et al., 2023, p. 270)

In their article, Pinney et al. further differentiate different definitions of gender and how they are commonly used in IAS research literature. The authors differentiate between gender identity and gender expression, and break these notions down further as follows:

> "Gender identity typically refers to one's own internal understanding of gender and self-identification. Gender expression refers to how one presents one's own gender and wants to be seen by the world. These both can fit into binary notions of gender, but can also be expansive and encompass a constellation of different identifications and notions of what self-expression can entail. More- over, gender expression can be broken up both internally (how one is expressing one's gender and feels about it to themselves) and how others perceive that individual's gender (perceived gender expression)." (Pinney et al., 2023, p. 270f.)

For the purposes of their article, Pinney et al. "focus on discussing gender, given that technological artifacts and systems typically discuss social constructions of gender as datafied by informational systems." (Pinney et al., 2023, p. 271) 271 They furthermore do not distinguish between gender identity and gender identity due to the many sources from which the gender variable stems from in the literature they analyze. They also note that these categories are used interchangeably in these contexts, anyway.

Similarly, the gender term used in this thesis' content analysis will be used without distinguishing between gender expression and gender identity, and there will be no further distinction between sex and gender, which has been previously discussed in the subchapter on Judith Butler's gender definition. The reasoning behind this is also similar: the gender data used in the text corpus is obtained in many different ways, and furthermore, this paper will mainly examine whether the gender category is defined in a binary structure, and whether there exists an effort to implement a more inclusive gender variable in the future.

Pinney et al. collected papers from 2017-2021 using the ACM Digital Library search and created a codebook based on their research questions (Pinney et al., 2023, p. 271). The categories in the codebook captured whether there is a gender variable present, the primary referent of the gender variable (who is the gender variable being attributed to?), which and how

many values could be attributed to the gender category, as well as whether the gender value is self-determined, or whether it is inferred by a third party. They furthermore captured whether the paper is about fairness or bias and what the general goal of the paper is. Lastly, they made a note on whether an author (group) acknowledged that gender was non-binary while still operationalizing it in a binary manner (Pinney et al., 2023, p. 272).

Several results were noteworthy: First, they found that most of the literature they reviewed relied on a binary notion of gender and that none of the papers "successfully affirmed or accounted for non-binary gender identities." (Pinney et al., 2023, p. 275) Secondly, there was a visible increase in awareness of the issue of gender fairness in this research community. More effort was found in papers with the goal of auditing system behavior, whereas papers with the goal of gender personalization or gender prediction failed "to properly analyze the implications of their findings or model behavior in reference to gender bias and fairness." (Pinney et al., 2023, p. 275) The authors suggest that "it may be the case that these two types of goals are antagonistic or fundamentally at odds with fairness and ethics," (ibid.) referencing work by Scheuerman et al. as well as Keyes (Scheuerman et al., 2019) (Keyes, 2018). Third, Pinney et al. found that the gender variable was most frequently used as input in the context of a user study or survey. They found this to be an encouraging result since there seems to be an effort to study how differently gendered users respond to a system, however, they found it worrying how frequently "systems attempt to personalize results based on gender. This itself makes major assumptions about what individuals may prefer, based on a gender variable, rather than on user preferences" (Pinney et al., 2023, p. 275). Lastly, the study showed that gender self-identification was the most frequent mode of determination, which they have found to be the most ethical way of obtaining gender data, since third-party annotators may act on and perpetuate gender stereotypes and misgender individuals. Nonetheless, self-identification does not guarantee the absence of misgendering, since data collected with solely binary options will inevitably misgender individuals anyway.

In conclusion, the authors of *Much Ado About Gender* recommend using an inclusive concept of gender, urging researchers to document their gender label collection with precision, and to consider more gender diversity in their data samples. They also recommend using the gender variable in computer science research to audit system for fairness. This, however, has its exceptions, since

> "work that aims to improve fairness but only does so within a binary gender construct, for example, may reinforce discrimination against non-binary people. Moreover, audits of system behavior that infers gender on individuals may reproduce harm by guaranteeing that a system works only for individuals who conform to stereotypical gender presentations or expressions." (Pinney et al., 2023, p. 276)

Furthermore, the usage of gender variables for purposes of personalization and prediction are not recommended. This is due to the fact that gender personalization, even in the context of cold-start scenarios where little to no user data is available, can lead to a system learning stereotypes from gender stereotype assumptions as opposed to organic user interaction data, further enforcing said stereotypes even if they do not accurately model user behavior (Pinney et al., 2023, p. 276). Gender prediction has previously been criticized by Os Keyes in *The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition* and Pinney et al. agree on the notion that gender 'recognition' or prediction models (whether it be visual, textual, or other types of recognition) are inherently based on perpetuating harmful stereotypes and determining gender "based on data instances which bear no relationship to gender, and will most likely misrepresent individuals who are transgender or gender non-conforming" (Pinney et al., 2023, p. 275). On this note, Pinney et al. urge scientists to consider the "danger in collecting demographic information in and of itself" (Pinney et al., 2023, p. 277) , making a similar case for the dangers of datafication as has been made previously by Keyes, Hoffmann, and others.

2.6. Bias Mitigation Strategies

There exist various categories of bias mitigation strategies in recommender systems. This thesis will follow the convention of categorizing these strategies according to different machine learning pipelines: pre-, in-, and post-processing methods. Three papers out of the corpus have been chosen to briefly explain each one of those methods.

2.6.1. Pre-processing

Nandy et al. claims that "[p]re-processing modifies training data to reduce potential sources of bias, often by removing features that are correlated with protected attributes" (Nandy et al., 2022, p. 715). Earlier sections on statistical parity and equalized odds and equalized opportunity have already discussed why the removal of those features may lead to suppression and why, in some cases, this might be problematic. This does not mean, however, that all pre-processing techniques aim to suppress sensitive data. Mehrabi et al. state more generally that "[p]re-processing techniques try to transform the data so the underlying discrimination is removed […]. If the algorithm is allowed to modify the training data, then pre-processing can

be used“ (Mehrabi et al., 2021, p. 14). According to Caton & Haas, pre-processing techniques are mainly occupied with altering distributions tied to protected attributes:

> “Pre-processing approaches recognize that often an issue is the data itself, and the distributions of specific sensitive or protected variables are biased, discriminatory, and/or imbalanced. Thus, pre-processing approaches tend to alter the sample distributions of protected variables, or more generally perform specific transformations on the data with the aim to remove discrimination from the training data […]. The main idea here is to train a model on a ‚repaired‘ data set. Pre-processing is argued as the most flexible part of the data science pipeline, as it makes no assumptions with respect to the choice of subsequently applied modeling technique […].” (Caton & Haas, 2023, p. 5)

One famous example of bias in data comes from the field of Natural Language Processing. In order to read, retrieve, or generate text, NLP models often rely on accurate word representations which have meanings assigned to them, features attributed to them, as well as relationships with each other. These word embeddings are, however, far from being unbiased, as has been discussed by Buonocore in *Man is to Doctor as Woman is to Nurse: the Gender Bias of Word Embeddings,* in which he exemplified gendered bias via a mathematical equation as an analogy to the word representations of ‘man’, ‘woman’, ‘doctor’, and ‘nurse’: “doctor − man + woman = nurse” (Buonocore, 2019). Wang et al. have argued that in cases like word embeddings, adding a de-biasing step prior to the model training might be the best first step towards fair recommendation (Wang et al., 2023, p. 7) .

2.6.2. In-processing

Nandy et al. states that “[i]n-processing (also known as training-time) mitigation methods modify the model training objective to incorporate fairness, often by adding constraints or regularization penalties“ (Nandy et al., 2022, p. 715). Mehrabi et al. add onto this definition stating that they furthermore „try to modify and change state-of-the-art learning algorithms to remove discrimination during the model training process“ (Mehrabi et al., 2021, p. 14) and that this can be done whenever the strategy has access to the learning procedure of a machine learning itself, “either by incorporating changes into the objective function or imposing a constraint“. [ibid.] The dominant idea behind in-processing strategies according to Caton & Haas is that they

> “recognize that modeling techniques often become biased by dominant features, other distributional effects, or try to find a balance between multiple model objectives, for example having a model which is both accurate and fair. In-processing approaches tackle this by often incorporating one or more fairness metrics into the model optimization functions in a bid to converge towards a model parameterization that maximizes performance and fairness.“ (Caton & Haas, 2023, p. 5)

2.6.3. Post-processing

Post-processing techniques are described by Nandy et al. as methods which "transform model scores to ensure fairness according to a provided definition. Post-processing methods learn (protected-attribute-specific) transformations of model scores to achieve fairness objectives […]." (Nandy et al., 2022, p. 715) Mehrabi et al. furthermore state that this strategy allows for fairness to be upheld even in cases where the strategy cannot access the training data or the model itself:

> "If the algorithm can only treat the learned model as a black box without any ability to modify the training data or learning algorithm, then only post-processing can be used in which the labels assigned by the black-box model initially get reassigned based on a function during the post-processing phase […]." (Mehrabi et al., 2021)

Caton & Haas explain that post-processing techniques are based on the idea that it is the output, specifically, which is behaving unfairly with respect to a specific protected variable: "Thus, post-processing approaches tend to apply transformations to model output to improve prediction fairness." (Caton & Haas, 2023, p. 5) Similarly to pre-processing, the fact that only the predictions are needed in order to apply post-processing bias mitigation strategies allows for more flexibility. This makes it advantageous "for black-box scenarios where the entire ML pipeline is not exposed." (ibid.)

Lastly, Mehrabi et al. state that choosing the most beneficial bias mitigation strategy can be difficult. However, due to their flexibility, both pre- and post-methods can leverage open source machine learning libraries, since the model itself does not need to be changed. Nonetheless, these methods can come with another set of problems:

> "modification of the data and/or model output may have legal implications [..] and can mean models are less interpretable […], which may be at odds with current data protection legislation with respect to explainability. Only in-processing approaches can optimize notions of fairness during model training. Yet, this requires the optimization function to be either accessible, replaceable, and/or modifiable, which may not always be the case." (Mehrabi et al., 2021, p. 5)

3. Methods

The method of choice for this research paper analysis was Philipp Mayring's systematic content analysis. The content of this chapter is mainly based on the book *Qualitative content analysis: theoretical foundation, basic procedures and software solution,* which was published by Mayring in 2014 (Mayring, 2014).

3.1. Historical Context and Epistemological Consequences

Qualitative content analysis is a mixed-method approach, which needs to be contextualized in the scientific discourse regarding qualitative and quantitative approaches to research. Mayring describes the debate between these two paradigms as a "Science war" (Mayring, 2014, p. 6), explaining that qualitative methods have been mobilized by recent scientific developments which formulated the quantitative method of Randomized Controlled Trials as „the only valid scientific procedure." (ibid.) This prompted Norman Denzin to publish the qualitative manifesto *A call to arms* in 2010, which Mayring described as radically constructivist. Mayring states that "[i]f not coming from a position of radical constructivism (treating different positions as equivalent subjective constructions), this situation is extremely unsatisfying for experienced researchers and newcomers." (Mayring, 2014, p. 7)

The methodologist argues that there exist convergences in the strict contraposition of the hermeneutical position, which is a constructivist theory which "tries to understand the meaning of the text as interaction between the preconceptions of the reader and the intentions of the text producer" (ibid.), and the positivistic position, which "tries to measure, to record and to quantify overt aspects of the text" (Mayring, 2014, p. 8) while also claiming objectivity in the results of the analysis. The convergence of these two seemingly contrary belief systems is explained by Mayring as follows:

> "The social constructivist theory formulates the possibility of an agreement between different individual meaning constructions and allows by that the concept of a socially shared quasi-objective reality. Modern hermeneutical approaches try to formulate rules of interpretation. By this, the analysis gains objectivity. On the other hand, positivistic positions had been refined to post-positivism or critical rationalism […]. Here, only an approximation to reality, accompanied by critical efforts of researchers to falsify hypotheses, is held to be possible, representing again the notion of an agreement process in talking about reality instead of a naive copy of reality." (ibid.)

Additionally, Mayring attributes the reconciliation of these positions to the differentiation of the research process into different phases. Research can be split into the context of discovery, in which a researcher defines a research question and develops hypotheses, and a context of

justification, in which said hypotheses are tested (Hoyningen-Huene, 1987). These two phases have been later extended by a third period "of deriving praxis consequences from the research results (context of application)" (Mayring, 2014, p. 8). According to Mayring, social science research ought to follow the paradigm of critical theory in the context of discovery as well as application, whereas a moderate constructivist or post-positivist position should be applied to the context of justification in order "to guarantee scientific rigor." (ibid.) Mayring argues that this mixed-method approach has thereby not lead to the development of a new methodology per se, but is rather an application of multiple perspectives onto different parts of the research process, "mainly following a pragmatic theory of science (the methodology is adequate if it leads to the solution of the research question)." (ibid.)

3.2. Research process

Qualitative Content Analysis owes its methodological basis to Quantitative Content Analysis. As a mixed-method approach, its aim is to categorize passages of text into categories in a qualitative-interpretative manner by following quantitative rules of content analysis. In summary, Qualitative Content Analysis uses qualitative methods by interpreting text and assigning it into categories, as well as quantitative methods by analyzing the frequencies of said categories in a text corpus.

The next view sections will describe the research process inspired by Mayring's basic research steps and how they were applied in the context of this thesis (Mayring, 2014, p. 10f.).

The first step of research is meant to manifest questions relating to a topic, making the praxis of the research relevant, potentially leading to hypotheses. It is also used to formulate and explicate the researcher's standpoint. While quantitative methodology requires the formulation of hypotheses in a deductive manner, for qualitative works of research such as explorative studies (which formulates new categories out of existing material through inductive category development (Mayring, 2014, p. 12)) this requirement is usually softened since the initial formulation of hypotheses is not always possible. However, the inherent structure of researcher-subject-interaction behind qualitative processes generally implies that a standpoint has been formulated on behalf of the researcher, which Mayring also includes as being a form of hypothesis (Mayring, 2014, p. 10).

In the context of this thesis, research questions were formulated based on my background as a student of the *Artificial Intelligence* Bachelor program of the Johannes Kepler University in Linz, as well as through the first sighting of the material corpus. Skimming and accumulating research literature on the topic of fairness measures and bias mitigation in recommender systems was the first iteration which lead to the formulation of the following research questions:

RQ1: How are the concepts of fairness and gender currently defined in recommender system research?
RQ2: Which biases are most commonly treated and which bias mitigation strategies are currently the most common in the domain of recommender systems?

Iterative reformulation of research questions within the process of this qualitative content analysis was necessary in order to make the results of the analysis concrete and relevant to the research praxis.

3.2.1. Linking Research Questions to Theory

Next, a theoretical approach to answering the previously formulated research questions is applied. State of the art knowledge is needed to frame research questions and results within an existing scientific field. Mayring states that "every research process is influenced by (hidden or formulated) preconceptions and only by linking research to theory a scientific progress is possible." (Mayring, 2014, p. 11) The author states that this is especially relevant for the interpretation of results. Mayring summarizes that this research step includes the "the formulation of preconceptions in advance and the stepwise modification of those preconceptions in confrontation with the material" (ibid.).

In the context of this thesis, the theoretical background is based on the field of Critical Algorithm and Data studies, as it involves exploring technological systems such as recommender systems from a humanities background. Fundamentally, this work is based on works directly related to the intersection between Gender Studies and Information Technology. Literature has been extracted from my subjects in Gender Studies (MA) as well as Artificial Intelligence (Bsc), and by using both the ACM Digital Library and the u:search platform provided by the University of Vienna.

Fairness and bias in recommender systems have been mainly studied from a perspective of bias mitigation, however, not many prior works of research have dealt with a meta-analysis studying explicitly the *definition* of fairness used by these research papers in the first place. Questioning and analyzing the definition of such a fundamental concept as *fairness* and *bias* is essential to maintain the integrity of research dedicated to the dismantling of oppressive structures.

3.2.2. Definition of the Sample and Sampling Strategy

The sample of a study is the empirical basis of a research project. Sampling can be done through a variety of sampling strategies. Mayring notes that, most importantly, it is necessary for a researcher to avoid convenient or "ad-hoc-samples" (Mayring, 2014, p. 12) and that sampling size and sampling strategy must be justified and described accordingly.

In order to answer the previously formulated research questions, a corpus of recommender system research papers on the topic of bias mitigation had to be found. The choice of platforms for the the sampling for this study has been chosen according to diversity in results (as in: how much do the results overlap with previously tested platforms?) as well as accessibility (are the research papers readily downloadable in a readable format, i.e. as a PDF-document?) The final extraction of the material corpus has been done in September of 2023 on the Association for Computing Machinery Digital Library platform. A trial of several online platforms for research papers has been conducted prior to this extraction, and the ACM Digital Library resulted in by far more results than: the Digital Library of the Institute of Electrical and Electronics Engineers (IEEE), Google Scholar, arXiv, Semantic Scholar, SpringerLink, ScienceDirect, MicrosoftAcademic, DataBase systems and Logic Programming (DBLP), and JSTOR. Websites which publish papers on the topic but were mainly dedicated towards conferences (such as the Conference on Neural Information Processing Systems (NeurIPS), the International Conference on Machine Learning (ICML), or the Association for Computational Linguistics (ACL)) were not of interest as they had limited resources compared to the previously mentioned search engines.

The ACM Digital Library is a research networking platform and literature database containing a variety of full-text publications on the topic of computing and information technology. At the time of writing, the ACM full-text collection holds over 724,061 records and the extension to the ACM Guide to Computing Literature yields 3,624,095 records. The extraction was done by

extending the search to the ACM Guide to Computing Literature in order to obtain as many relevant results as possible.

i. Site-specific Filtering Criteria

In order to make the study feasible within the framework of a master thesis, filtering and expansion criteria had to be formulated in the case that the corpus was too small or too big. The main site-specific filtering option used on the ACM Digital Library was related to the accessibility of the document, which meant that only papers which were downloadable in PDF-format were considered as viable options. Furthermore, the option to only include research articles was used and other types of texts, such as short papers or abstracts, were excluded via site-specific filtering options. This was because research papers are more likely to explore the topic with sufficient depth.

*ii. **Keywords***

The following keywords were used to yield results relevant to the research questions:

- "recommender"
- "bias"
- "mitigation"
- "gender"
- "fairness"

Only papers which include all five keywords at any point in their content were considered. Singular terms were used in order to achieve the broadest possible range of results and in order to not miss any relevant items. The search for these keywords in the ACM Digital Library yielded 80 results in total.

iii. Further Filtering of the Corpus

During the initial sighting of these research papers, only those with more than ten (> 10) pages were considered as suitable options since shorter papers are less likely to have an in-depth examination of the topic. This excluded 12 more research papers, which left 68 documents in total.

In this iteration, papers were further excluded based on the content of their abstract. This step was necessary as the filtering via keywords did not guarantee that the papers in the search result were relevant with respect to the research questions. Reasons for filtering out papers from the corpus based on the abstract included:

- Paper's main subject of interest is not directly related to recommender systems.
- Paper examines topic of interest on a meta-level instead of directly mitigating bias with mitigation strategies and testing fairness with fairness measures.
    - In this case, the research paper was considered as a candidate for the theory corpus. An example for this is the paper *Much Ado About Gender* by Pinney et al. which examines gender definitions in IAS research papers.
- Paper is a duplicate of another paper.

Lastly, if the corpus included multiple papers by the same author or group of authors, the older document was excluded from the corpus. The reasoning behind this choice is that it is more likely for research papers to have similar definitions on fairness, bias and gender when written by the same author or group of authors.

After these filtering methods have been applied, the final corpus consisted of 24 papers.

Mayring states that Qualitative Content Analysis entails inductively developing categories out of the material or deductively formulating categories according to a knowledge base. For this study, a deductive structure was chosen because there already exists a lot of literature in this field containing structural suggestions regarding fairness metrics, gender categories, and bias mitigation pipelines. This process of categorization will be described in the following section.

3.2.3. Deductive Category Assignment

In *Qualitative Content Analysis*, Mayring describes three types of category assessment. Inductive category formation follows a logic of summarizing text into inductively created categories, which are summarized from the material itself (Mayring, 2014, p. 79). Explicational content analysis aims to do the reverse by enriching parts of text with the aid of additional material explaining them (Mayring, 2014, p. 88). Lastly, deductive (or structuring) category assignment is what Mayring describes as the "most central" (Mayring, 2014, p. 95) type of qualitative content analysis. The goal of this analysis is to extract a certain structure from a given material. For this, categories must be prepared in advance, based on theoretical

knowledge. Text components are then "addressed by the categories are then extracted from the material systematically." (ibid.) Mayring argues that the following points are especially important:

> "The fundamental structuring dimensions must be exactly determined. They must derive from the issue/statement of the problem concerned, and must be theoretically based. These structuring dimensions are then, as a rule, further subdivided, being resolved or split up into individual features or values. Subsequently, the dimensions and values are brought together to form a category system." (ibid.)

The type of category assignment chosen for this study was the deductive category assignment. This is because "deductive category assignment is the adequate procedure if there is relevant previous research" (Mayring, 2014, p. 97) and Critical Algorithm and Data Studies, specifically in the field of recommender systems, is already an established scientific field. This procedure is defined as a deductive process since the category system is established before the corpus has been sighted or coded. This category system is then revised in later steps. Further rephrasing or adding of categories can happen based on theoretical considerations, however, "categories are not developped out of the text material like in inductive category formation." (ibid.) The iterative revision of the corpus used by this study has, for example, lead to some categories being omitted or merged, which will be discussed at the end of this chapter.

Without further ado, the three stages of categorization are defined as:

1. Definition of categories: the precise determination of which category a specific piece of text belongs to.

Mayring bases this process on general psychology and its theories of categorization (Mayring, 2014, p. 37). In this field of study, the processes of learning, memory, and a person's mental representation of the world are analyzed. Mayring states that "[c]oncepts are mental representations of classes of things" (ibid.) and categories are those classes which help conceptualize the world. He further explains that the most precise approach towards categorization would be one that ideally uses three main theories of categorization according to George L. Murphy's *The Big Book of Concepts*: (Murphy, 2002)

- **The definitional theory** states that classification of objects should be based on a set of sufficient conditions for belonging to a specific category.
- **The prototype theory** states that for every category there exists a typical example which illustrates the category. Similar objects can then be categorized in the same class.

This approach, however, has its limitations since some exemplars of a certain categories may be less typical than others, which makes delineation difficult.

- **The decision bound theory** states that "categories are defined by their differences to neighbor categories" (Mayring, 2014, p. 38). According to this logic, an object can be categorized when it doesn't fulfil the conditions of a different, albeit similar, category. The limitation of this approach is that the aspect of mental representation of an object is undefined.

2. Anchoring examples: Specific passages in text aiding as typical examples to illustrate the category this piece of text belongs to. This part of Mayring's procedure directly implements the prototype theory.

3. Coding rules: In order to make assignment to categories unambiguous, coding rules aid in the delineation between categories. Similarly, this approach leans on Murphy's definition of the decision bound theory.

The following segment will now present the category system as well as describe its definitions and coding rules in detail.

3.3. Codebook

Fairness measure used

- Purpose of category: This category aims to specify which type of fairness measure was used in the paper, or which fairness measure is proposed as a solution by the paper. The focus lies on the fairness measure which is actually used and presented in the paper, as opposed to a fairness measure which might be merely mentioned in the 'discussion' or 'future work' part of the paper. This is to reduce complexity in the coding process as well as to stay focused on fairness measures which are addressed in sufficient depth.
- Further specifications:
    - Wherever possible, the differentiation between 'singular attributes only' (marked as 'singular') and 'combinations of attributes possible' (marked as 'multiple') has further been applied in order to assess where the tendency lies when it comes to the option of including intersectional identities in a paper's study on fairness measures.

- Wherever this information is available, an additional specification ‘U’, ‘P’, or ‘I’ is added to differentiate between user-, provider-, or item-side fairness, or a combination of those.

0. undefined

- Description: The fairness measure was not sufficiently described in the paper or its definition was omitted entirely.
- Example: The paper’s purpose is to introduce a set of biases and bias mitigation strategies and doesn’t mention any specific fairness measure(s).

1. Individual fairness

- Description: Individual fairness is the opposite of group fairness and generally defined via a similarity metric. It is seemingly independent of protected attributes. The use of this metric may or may not be due to a ‘fairness through unawareness’ approach, or due to the aim of maintaining fairness on an individual level via a problem-specific similarity metric (fairness through awareness).
- Coding rules:
  - In order to be a part of this category, a paper must not use any form of protected attribute for their fairness measure, as well as no placeholder attributes such as ‘advantaged’ vs. ‘disadvantaged’ groups.
  - Additionally, any fairness measure based on group splitting, even if that split not based on an attribute protected under law (for example, a random group attribute may be hair color) will not be considered for this category.
- Example: The fairness measure proposed by this paper is a new type of individual fairness, where no specific attributes are mentioned: “To address this issue, in this paper, first, we define a new notion of individual fairness from the perspective of items, namely (α, β)-fairness” (Wang & Wang, 2022, p. 1)

2. Group fairness

- Description: The paper’s study explicitly mentions protected attributes such as gender, race, age, or other attributes which are considered as a protected attribute. Alternatively, the paper mentions placeholder-terms such as ‘protected attribute’ or ‘sensitive

attribute'. Alternatively, the paper mentions any type of group-split in order to assess the fairness of the recommender system.

- Coding rule: As previously mentioned, protected attributes are well-defined attributes under law. However, the attributes 'disadvantaged' or 'advantaged' groups will also be counted as protected attributes. This is due to the fact that many papers in this context use the differentiation of 'advantaged vs. disadvantaged groups' or 'majority vs. minority groups' as placeholders for actual protected attributes. Those cases will not be further distinguished from another as it does not matter for the purposes of this study.
- Examples:
    - In *Comprehensive Fair Meta-learned Recommender System*, the term 'sensitive attributes' is specifically mentioned in the Discussion-part of the paper: "In the future, we would like to design fairness metrics for multi-class sensitive attributes and explore the fairness issues within the combination of multiple sensitive attributes." (Wei & He, 2022, p. 1997)
    - In *Measuring and Mitigating Item Under-Recommendation Bias in Personalized Ranking Systems*, a framework was proposed which aims to achieve fairness across both individual and group fairness, calling it 'subgroup fairness'. Since a group split (as well as the use of protected attributes) is still involved, this fairness measure will be categorized as a group fairness measure:

> "Individual Fairness requires that systems should give similar predictions to individual users and content generators with similar characteristics, regardless of their differences in protected sensitive attributes, such as gender, ethnicity, and popularity […]. Group Fairness concept focuses on the potential biases against sensitive groups or communities and emphasizes that all groups should be treated equally […]. Subgroup Fairness combines the features of both fairness concepts above and measures whether a fairness constraint holds over a large set of subgroups […]." (Zhu et al., 2020, p. 237)

3. Agnostic / multiple

- Description: The fairness measure is explicitly described as flexible or agnostic, meaning that both individual and group fairness achieved in this context. Some researchers might choose this method in order to show how a new model of a recommender system can be proven to make decisions fairly based on multiple fairness measures. Alternatively, this paper provides an introduction into a multitude of fairness measures, without focus on either one of them.

- Coding rule: This category is similar to the ‘undefined’ category, since some papers are ambiguous when it comes to their choice of fairness measure. However, the main difference lies in whether the paper does not mention a specific fairness measure because its focus lies on a different type of method, such as bias mitigation (‘undefined’ fairness measure), or whether it mentions and perhaps describes multiple fairness measures or groups of fairness measures in a non-hierarchical fashion (‘flexible/agnostic’ fairness measure). The latter is especially obvious in papers which use both individual and group fairness measures.
- Examples:
    - The paper does not specify specific fairness measures and instead uses broad terms such as ‘individual’ and/or ‘group fairness’ which implies the possibility of multiple fairness measures which measure fairness based on user attributes, for example: “Additionally, we devise the group fairness and individual fairness criteria with regard to recommendation performance and explanation diversity.“ (Fu et al., 2020, p. 70)
    - The paper uses a cluster or a family of fairness measures, such as: „We formalize and compare a family of JME-fairness measures that deal with different types of systemic biases in content exposure from recommender systems.“ (Wu et al., 2022, p. 704)

- Subcategories:

PA. Statistical Parity

- Description: The fairness measure is (at least in part) based on statistical (or demographic) parity, which states that the probability output of the recommender system is independent of the protected attribute.
- Example: The paper states explicitly that it uses (some form of) statistical/demographic parity, for example: “our analysis of long-term dynamics of fairness interventions in connection recommender systems begins by assuming demographic parity of exposure as fairness measure.” (Akpinar et al., 2022, p. 23)

OD. Equalized Odds

- Description: The fairness measure is (at least in part) based on equalized odds, which states that equality must be achieved among only between individuals whose outcomes are similar (i.e. who have the same probability to be hired, or not hired).
- Example: The paper states explicitly that it uses (some form of) equalized odds as a fairness measure, for example: "We propose scalable methods for achieving equality of opportunity and equalized odds in rankings in the presence of position bias, which commonly plagues data generated from recommender systems." (Nandy et al., 2022, p. 1)

OP. Equalized Opportunity

- Description: The fairness measure is (at least in part) based on equalized opportunity, which states that only qualified members of each subgroup are treated with the same probability for a positive outcome (i.e. the probability to be hired).
- Example: See example for category 'OD. Equalized Odds' which includes both metrics.

DI. Disparate Impact

- Description: The fairness measure is (at least in part) based on disparate impact, or disparate (mis)treatment, which states the probability of an outcome should not be different when it comes to the protected attribute of the user, and that features correlating with the protected attributes must also be taken into account, since they might unintentionally lead to a disparity of outcomes.
- Example: The paper states explicitly that it uses (some form of) disparate impact as a fairness measure, for example: "With the pair-wise approach we employed in this work and the (un)fairness metric we will present, female providers receive 2.9% of the total item relevance (and 2.8% of exposure), being affected by the disparate impact." (Boratto et al., 2021, p. 424)

CF. Counterfactual Fairness

- Description: The fairness measure is (at least in part) based on counterfactual fairness, counterfactual inference, or any type of counterfactual scenario. This notion of fairness captures the idea that the fairness of an outcome should be independent of the protected attribute, and that this can be achieved by creating a scenario in which the protected attribute is exchanged for a different one. Fairness, according to this logic, is when the decision remains the same within the counterfactual and the actual world.
- Example: The paper states explicitly that it uses (some form of) counterfactual fairness as a fairness measure, for example: "As such, we develop a causality-enhanced User-Controllable Inference (UCI) framework, which can quickly revise the recommendations based on user controls in the inference stage and utilize counterfactual inference to mitigate the effect of out-of-date user representations." (Wang et al., 2022, p. 1251)

CA. Calibration Fairness

- Description: The fairness measure is (at least in part) based on calibration fairness, which aims to combat popularity bias, stating that a user's recommendations should be related to their interests, which in turn are based on their ratings. Calibration fairness aims to minimize the gap of how popularity bias impacts different users.
- Example: The paper states explicitly that it uses (some form of) calibration fairness as a fairness measure, for example: "In addition, we investigate the underlying relation between user-oriented fairness and user-oriented evaluation of popularity bias, known as popularity bias calibration fairness [2, 32], i.e., the degree to which popularity bias impacts users with different levels of interest in popular items in their profiles." (Rahmani et al., 2022, p. 2756)

OT. Other

- Description: The fairness measure is based on another fairness measure which is a function that was not mentioned in the category system so far. The papers in this category are the ones which either provide a completely novel fairness measure, or the ones which use a domain-specific fairness measure.

- Example: In *Protected attribute guided representation learning for bias mitigation in limited data*, the authors test their models via a bias score function which "will denote whether the model prediction is biased towards any protected attribute value." (Mazumder & Singh, 2022, p. 10)

Framing of source of discrimination

- Purpose of category: This category aims to capture the way in which discrimination is framed within the research paper in order to ascertain which discursive frames are prioritized over others.

0. undefined

- Description: The paper does not frame discrimination in any discernable or specific way. Alternatively, the paper does not discuss discrimination to a sufficient degree or omits mentioning discrimination entirely.
- Examples:
  - The term 'discrimination' does not appear in the text.
  - Bias in this paper is framed as something to be mitigated, but its source is not discussed in sufficient depth.

1. lack of inclusion

- Description: The paper frames the existence of discrimination as the result of exclusive data pools.
- Example: The introductory section contextualizes the bias mitigation strategy, which is a method aimed at combatting bias due to sparse or unbalanced data. In this section, it framed as if bias is the result of said faulty data as opposed to systemic power relations or other reasons. The bias mitigation strategy therefore entails duplicating or normalizing data in a way to combat the unbalance.

2. bad actors

- Description: The paper blames the source of discrimination on 'bad actors' in a similar fashion as Anna Hoffmann has described it in this theory subchapter.

- Example: The paper explicitly ascribes the cause of bias to be based on ‘bad’ algorithms and/or ‘bad’ data, for example: “System-caused bias, which we can also name model-intrinsic bias, occurs during the training and deployment of recommender systems. It is caused by unfair data or inducted by the model itself.” (Masri & Assi, 2014, p. 2711)

3. unconscious bias

- Description: Discrimination is framed as being the result of an unconscious or unintentional cognitive process.
- Example: Terms such as ‘unintentional’ are used when describing the source of bias, for example: “Prior work showed that certain minority groups of providers, characterized by a common sensitive attribute (e.g., gender or race), are being disproportionately affected by indirect and unintentional discrimination" (Boratto et al., 2021, p. 421)

4. issue of distribution

- Description: Discrimination is framed as being the result of an unbalance in the distribution of goods, rights, opportunities, and resources.
- Example: The authors of the paper ascribe the source of bias to be an unbalanced distribution of goods, rights, opportunities, and/or resources, and addressing them entails a form of redistribution.

5. systemic issue

- Description: Discrimination is explicitly described or defined as a systemic issue, rooted in complex socio-economic dynamics.
- Example: The paper mentions systems of power as being the source of bias in data, for example: “The problem of how individual or groups of items may be systemically under or over exposed to groups of users, or even all users, has received relatively less attention.“ (Wu et al., 2022, p. 703)

6. superiority of computer systems over humans

- Description: Discrimination is framed as being due to the irrationality of humans, who are deemed to be naturally more biased, whereas computer systems are inherently less biased and more ‘objective’.
- Example: Bias is largely framed as human bias, for example: „human bias refers to the systematic deviations of human behavior from the predictions of rational normative models. In contrast to the assumptions of many simulated user models, people are boundedly rational and their decisions are often affected by a series of biases and mental shortcuts“ (Liu, 2023, p. 236).

Gender definition used / proposed

- This category aims to capture if and how gender is mentioned and framed in the context of the given research paper.
- Further specification: Wherever applicable, it is differentiated between definitions of gender used and definitions of gender proposed by the paper. This is order to ascertain whether there are papers who make this distinction, for example by proposing to use gender as a non-binary variable despite using it as a binary variable in their experiment.

0. undefined:

- Description: This paper doesn't mention or define gender (to a notable degree). Alternatively, this category is not applicable.
- Examples:
    - The paper does not use gender as a variable, therefore there is no definition of gender.
    - The paper mentions gender as a variable but does not provide any further specifications.

1. gender is addressed

- Description: This paper explicitly or implicitly addresses gender as a variable and describes its use-case to a sufficient degree.
- Subcategories:

1.1. binary

- Description: Gender is defined as a binary variable.
- Example: Gender is defined as a binary variable, which may or may not be ‘male’ and ‘female’: “To construct user groups, we split MovieLens users into two groups by gender.” (Liu et al., 2023, p. 16)
- Coding rule: To delineate this category from the ‘undefined’ category, any set of binary variables count as a binary definition of gender, even if they are not ‘male’ and ‘female’. For example, in *Fairness metrics and bias mitigation strategies for rating predictions*, gender is defined as one possible variable substituting the ‘advantaged’ and ‘disadvantaged’ groups (Ashokan & Haas, 2021).

1.2. non-binary

- Description: Gender is defined as a non-binary variable.
- Example: Gender is defined in a multi-variable fashion, which may or may not be ‘male’, ‘female’ and ‘other’/’do not want to disclose’/’non-binary’ and so on. For example:

> “Our simulation framework allowed us to evaluate the algorithms over attributes with up to 10 values (e.g., <gender, age group> which could assume 9 values with three gender values (male, female, and other/unknown) and three age groups), and also study the effect of varying the number of possible attribute values.“ (Geyik et al., 2019, p. 2225)

Type of bias treated

- Purpose of category: This category aims to capture which type of bias was treated by the given research paper.

0. undefined

- Description: The paper does not treat bias or does not define the type of bias treated to sufficient depth. Alternatively, the paper provides novel fairness testing approaches which aim to treat multiple forms of bias (often claimed to be ‘general’ or ‘universal’) in order to adapt them to a new system. Similarly, when a research paper is more like a summary of biases which can appear in specific contexts, it will also fall into this category.
- Examples:

- The term ‘bias’ does not appear in the text.
- The paper focusses on topics other than bias mitigation, such as fairness measures, and therefore does not provide a description or definition of bias.
- The paper’s provides descriptions of multiple bias types, but its main focus lies on proposing a new approach towards fairness testing for a new system, such as deep recommender systems: “Fairness testing of deep learning models is an emerging research area in software engineering that aims to expose multiple kinds of fairness issues of a deep learning model, e.g., individual discrimination […], group disparity […], etc. […] In this work, we propose FairRec, a novel unified fairness testing framework specifically designed for DRSs to address the above challenges.“ (Guo et al., 2023, p. 311)
- The paper introduces multiple biases which can appear in a specific type of recommender system, such as in *Biases in scholarly recommender systems: impact, prevalence, and mitigation* (Färber et al., 2023).

1. Data Generation bias

- Description: The type of bias which is treated in the given paper stems from issues during the data generation process, before any model building has occurred.
- Coding rule: Multiple bias types may be mitigated in one research paper. This means more than one subcategory may be denoted for each document.
- Subcategories:

1.1. Historical bias

- Description: The paper aims to mitigate bias caused by systemic societal power structures. This means bias which occurs “even if data is perfectly measured and sampled, if the world *as it is* or *was* leads to a model that produces harmful outcomes.“ (Suresh & Guttag, 2021, p. 4)
- Example: The authors explicitly or implicitly state that the bias to mitigate is, first and foremost, caused by socio-economic issues embedded in history. For example: “Specifically, we consider group attributes for both types of stake-holders to identify and mitigate fairness concerns that go beyond individual users and items towards more systemic biases in recommendation.” (Wu et al., 2022, p. 703)

1.2. Representation bias

- Description: The paper aims to mitigate representation bias, such as training data being highly skewed and under-representing a specific demographic, subsequently failing to generalize the model for a subset of the population.
- Coding rule: This subcategory also includes bias that is created due to data sparsity due to inactive users, which typically results in a lack of recommendation diversity. Generally, any type of bias (even if not explicitly referred to as 'representation bias' in the research paper) which is caused by missing training data, for any reason, will be put into this category.
- Examples:

    - The authors explicitly or implicitly state that the bias to be mitigated in their research paper is bias caused by a skewed representation of certain user groups in the dataset. For example: "This allows for a flexible creation of synthetic data which represents different biases. Specifically, we consider the following scenario. The user population is skewed in the sense that 40% belong to type W, 10% to type WS, 40% to type MS, and 10% to type M." (Ashokan & Haas, 2021, p. 12)
    - The authors explain that the underlying cause for the bias is missing training data, for example: "We show that inactive users may be more susceptible to receiving unsatisfactory recommendations, due to insufficient training data for the inactive users, and that their recommendations may be biased by the training records of more active users, due to the nature of collaborative filtering, which leads to an unfair treatment by the system." (Fu et al., 2020, p. 69) In this case, inactive users can be considered as the group which is under-represented in the dataset.

1.3. Simpson's Paradox

- Description: The paper aims to mitigate bias caused by the Simpson's Paradox, which occurs when a subset of a dataset shows different statistical trends than the overall dataset.
- Example: The authors explicitly or implicitly state that the bias to mitigate is caused by the Simpson's Paradox. For example: "Experiments are conducted on two types

of real-world datasets—traditional and randomized trial data—and results show that our framework can im- prove the recommendation performance and reduce the Simpson's paradox problem of many CF algorithms." (Xu et al., 2023, p. 235)

2. Model Building bias

- Description: The type of bias which is treated in the given paper stems from issues during the model building process.
- Subcategories:

2.1. Popularity bias

- Description: The paper aims to mitigate bias caused by the overexposure of popular items.
- Example: The authors explicitly or implicitly state that the bias to mitigate is caused by popularity bias. For example: "To address this issue, in this paper, first, we define a new notion of individual fairness from the perspective of items, namely (α, β)- fairness, to deal with item popularity bias in recommendations." (Wang & Wang, 2022, p. 117)

2.2. Algorithmic bias

- Description: The paper aims to mitigate bias caused by algorithmic bias, which occurs when bias is added by the algorithm of a recommender system while the input data is assumed to be unbiased.
- Coding rule: For the purposes of this study, algorithmic bias will mainly be differentiated from algorithmic bias in two ways:
    - Demographic bias is treated by those papers who explicitly mention ways in which algorithmic bias affects different demographic groups in different ways, i.e. whose focus is to maintain fairness amongst all demographics.
    - Algorithmic bias is treated by those papers who focus more on model-specific problems than on effects on different demographic groups, additionally to stating that the training data is assumed to be unbiased.

- Example: The authors explicitly or implicitly state that the bias to mitigate is caused by algorithmic bias. For example: “This paper focuses on the detection and mitigation of algorithmic bias defined through predictive parity for large-scale AI systems.“ (DiCiccio et al., 2023, p. 1)

2.3. Demographic bias

- Description: The paper aims to mitigate bias caused by demographic bias, where a certain demographic in the population is treated differently by the algorithm of the recommender system
- Coding rule: If it is unclear if the bias type of a given paper belongs to ‘demographic bias’ and ‘representation bias’, the category is chosen by which term the authors of that paper explicitly choose to describe the bias they are trying to mitigate. Otherwise, further investigation needs to be made in order to ascertain whether the bias in this context is caused by a skewed dataset or by the algorithm.
- Examples:
  - The authors explicitly or implicitly state that the bias to mitigate is caused by demographic bias, exemplifying it by mentioning how a specific user group might be affected differently by the algorithm, or by arguing that the bias affects minority groups in particular:

    > “If these relevances are biased against the minority group, the recommender system is unfairly giving minority items less chance of being ranked high. Given its connection with the final ranking, relevance is thus an *internal algorithmic asset* to be allocated to provider groups, and not just a property of user–item pairs to be estimated.“ (Boratto et al., 2021, p. 422)

  - The paper addresses the cold start problem in recommendation. This is usually mitigated by using user attributes (such as sensitive attributes) in order to make baseline predictions. However, these predictions typically result in stereotyping. Therefore, the type of bias being mitigated in this context would be demographic bias.

3. Deployment and User Interaction

- Description: The type of bias which is treated in the given paper stems from issues during the deployment and user interaction phase of the recommender system building process.

- Subcategories:

3.1. Social bias

- Description: The paper aims to mitigate bias caused by social bias, more specifically, by users being influenced by the judgement of others.
- Example: The authors explicitly or implicitly state that the bias to mitigate is caused by social bias. For example, one paper aims to mitigate biased behavior influenced by self-efficacy: “In terms of overcoming gender bias in career decisions, the most commonly cited interventions include the availability of role models in the same social circle, especially same sex role models for women […]. Encouragement from friends and family […] also improves self-efficacy.“ (Wang et al., 2023, p. 5) The authors furthermore aimed at mitigating social bias by using surveys: “We conducted a follow-up survey to gain additional insights into the efectiveness of various design options that can help participants to overcome their own biases“ (Wang et al., 2023, p. 1).

3.2. Interaction bias

- Description: The paper aims to mitigate bias caused by interaction bias, which occurs due to interaction differences between users and a recommender system.
- Coding rule: Although similar, interaction bias will not be merged with the category ‘Filter bubbles’. Interaction bias encompasses a variety of cultural and cognitive biases and are connected to the direct interaction of a user with a platform (clicking, hovering, lingering, etc.). Filter bubbles is a more specific (albeit different) category of biases which highlights the personalization aspect of recommender algorithms.
- Example: The authors explicitly or implicitly state that the bias to mitigate is caused by interaction bias. For example: “Within the framework, we also analyze the interactions between human and system biases in search and recommendation episodes.“ (Liu, 2023, p. 236)

3.3. Ranking bias

- Description: The paper aims to mitigate bias caused by ranking bias, which is caused by higher-ranked items being clicked on more often than others.

- Examples:

    - The authors explicitly or implicitly state that the bias to mitigate is caused by ranking bias, mentioning exposure and/or visibility as important factors in the context of ranking. For example: “We provide measures of fairness that allow us to assess the exposure of users of different groups in the ranking and propose a re-ranking algorithm to guarantee a fair exposure.“ (Boratto et al., 2022, p. 839)
    - The authors use the term ‘position bias’ referring to the position of the item in a rank: “We formally extend the definitions of equality of opportunity and equalized odds from the binary classification […] to the ranking context, provide a causal interpretation of the equalized odds condition […], and provide scalable post-processing techniques for mitigation of bias identified through these definitions, which has not been addressed by previous literature on fairness in ranking. We also explicitly handle the position bias issue and suggest simple mechanisms for controlling the fairness versus performance trade-off.” (Nandy et al., 2022, p. 716)

3.4. Filter bubbles

- Description: The paper aims to mitigate bias caused by filter bubbles, i.e. the intellectual and social fragmentation and isolation of users on internet platforms.
- Example: The authors explicitly or implicitly state that the bias to mitigate is caused by filter bubbles, or frame the underlying issue as being due to personalized algorithms. For example: “This work proposes a new recommender prototype called User-Controllable Recommender System (UCRS), which enables users to actively control the mitigation of filter bubbles.“ (Wang et al., 2022, p. 1251)

Bias mitigation method type

- Purpose of category: This category aims to capture which type of method was used, proposed, or investigated, in order to mitigate a specific type of bias.

0. Undefined

- Description: The paper does not define a specific bias mitigation method to sufficient depth.
- Example: The paper focusses on topics other than bias mitigation, such as fairness metrics, and therefore does not provide a bias mitigation method.

1. Pre-processing

- Description: The paper's used or proposed bias mitigation method is a pre-processing method, which transforms the training data and therefore aims to remove the underlying cause of discrimination.
- Example: The authors state explicitly or implicitly that their bias mitigation method is a pre-processing method or that training data has been altered in a pre-processing step before being passed to the recommender system. For example, Wang et al. use a pre-processing step by altering the word embedding of a dataset: "To gender-debias the recommendation, we add a de-biasing step prior to applying logistic regression. Our debiasing approach adapts the work on attenuating bias in word vectors […]. Since traditional word embeddings are usually trained on massive text data, they inherit some of the human racial and gender biases from the data" (Wang et al., 2023, p. 7).

2. In-processing

- Description: The paper's used or proposed bias mitigation method is an in-processing method, which either adds to the existing algorithm or changes the algorithm of the recommender system itself in order to produce a fairer outcome.
- Example: The authors state explicitly or implicitly that their bias mitigation method is an in-processing method, which usually means that an algorithm or a set of algorithms were proposed in order to enhance an existing recommender system. Any type of proposed machine learning model generally points towards an in-processing procedure, for example: "Next we propose a novel debiased personalized ranking model incorporating adversarial learning to augment the proposed bias metrics." (Zhu et al., 2020, p. 458)

3. Post-processing

- Description: The paper's used or proposed bias mitigation method is a post-processing method, which is implemented after the model has already been trained.
- Example: The authors state explicitly or implicitly that their bias mitigation method is a post-processing method, for example by involving some sort of re-ranking strategy:

> "We observe disparities that favor the most represented groups. We overcome these phenomena by introducing equity with a re-ranking approach that regulates the share of recommendations given to the items produced in a continent (*visibility*) and the positions in which items are ranked in the recommendation list (*exposure*), with a negligible loss in effectiveness, thus controlling fairness of providers coming from different continents." (Gómez, Boratto, et al., 2022, p. 1)

4. Multiple bias mitigation methods used

- Description: The paper's uses or proposes a combination of pre-, in-, and/or post-processing methods.
- Examples:
    - The authors introduce and explain a multitude of bias mitigation methods.
    - The authors propose a bias mitigation method which includes the possibility to implement it in an in-, and post-processing fashion, depending on use-case.

Omitted and merged categories

The iterative revision of the deductive category system has lead to some categories being merged or omitted in order to reduce complexity as well as the amount of redundant information conveyed by the category system. For example, in the category of fairness measures, the categories 'DM. Disparate Mistreatment' and 'DI. Disparate Impact' have been merged into the latter term in order to reduce complexity. For the purposes of this study, the further differentiation between those two fairness concepts is not necessary, as they are based on the same logic: features correlating with protected attributes must be taken into account in order to prevent the unintentional disparity of outcomes. Moreover, several values for the equality measure category have been deleted, such as treatment equality, general entropy index, as well as in the bias type category: measurement bias, population bias, sampling bias, evaluation bias, aggregation bias, omitted variable bias, temporal bias, behavioral bias, content production bias, linking bias, emergent bias, observer bias, and presentation bias. These

categories have been omitted because they did not come up in the corpus, and including them would be arguably too complex. In conclusion, there exist bias types and fairness measures which have not been included in this study because the aim of this study is to address the types of fairness measure and bias categories which *did* come up. Although this will not be discussed in detail, it is important to acknowledge that the fact that these biases did not appear in the corpus still yields scientific insight.

4. Results

This chapter will summarize the results of the Deductive Qualitative Content Analysis conducted on the corpus containing 24 research papers on bias mitigation and fairness metrics for recommender systems.

4.1. Summary of Main Narrative

First, a summary of the main narrative structure of this master thesis will be provided. In the theory chapter, multiple terms and concepts were introduced and contextualized in the field of recommender systems, such as fairness, bias, and gender, as well as fairness measures, bias mitigation strategies, and stakeholders. These terms will be reiterated shortly and brought into context with the research questions of this thesis, which are:

RQ1: How are the concepts of fairness and gender currently defined in recommender system research?
RQ2: Which biases are most commonly treated and which bias mitigation strategies are currently the most common in the domain of recommender systems?

4.1.1. Fairness, Discrimination, and Gender

Anna Lauren Hoffmann has described in *Where Fairness Fails* multiple tendencies in the discourse around fairness and antidiscrimination, which appear to be even more prevalent in the field of information technology. Taking these tendencies into account, an ideal fairness framework would be one which…

1. …situates fairness and discrimination in a broader societal context and is aware of the structural role of technology. This would combat the common framing of ‘bad actors’ being responsible for unfairness, which blames either individual people, datasets, or algorithms for a problem which is bigger and more complex than these individual issues.

2. …is intersectional and aware of the conditions which make privilege and advantage possible in the first place. This would address the single-axis thinking and centering of disadvantage of current antidiscrimination discourse.
3. …is aware of and actively concerned with the structuring role of social attitudes and human dignity. This would attend to the common issue of framing discrimination as merely a distribution issue.

Moreover, a gender framework which adheres to the suggestions which Os Keyes et al. made in *You Keep Using That Word* would be one that decomposes specifically those gendered aspects which are relevant to their work, and then finds appropriate ways to measure those aspects. This could include one or more of the following conditions:

- The researchers define their concept of gender and any participants of the study, their concept of gender is taken into consideration as well.
- The researchers define the role which gender plays in their research.
- The researchers explain who and what is left out and how they are accountable for their work.
- The researchers inquire into concepts such as masculinity, which may have been systematically understood as a 'given' and thereby ignored.

As a prerequisite, it must be noted that Keyes et al. follow Judith Butler's notion that the regulation of gender as a binary structure is based on compulsory heterosexuality. Ignorance is posed as not just an absence of knowledge, but rather as "an active, socially shaped and cultivated phenomenon." (39:9) On that note, those who are left out by the incessant construction of binary gender structures are those who deal with the consequences of these structures, more than anyone else. Similarly, Pinney et al. have stated in *Much Ado About Gender* that even works of research dedicated to the practice of auditing systems for fairness ought to have a broader gender concept, since "audits of system behavior that infers gender on individuals may reproduce harm by guaranteeing that a system works only for individuals who conform to stereotypical gender presentations or expressions." (Pinney et al., 2023, p. 276)

Moreover, the categorization of gender into reductive labels can be seen as instances of data violence, as has been described by Hoffmann. Lastly, the way in which discrimination is framed matters on a discursive level. Hoffmann also coined the terms discursive violence, which captures the notion of why ideas about gender, fairness, and discrimination matter, and

why the discourse around it is anything but harmless. Violence can be discharged not just materially but discursively as well, perpetuating norms and ways of being, making the condition of *othering* possible.

4.1.2. Biases and Bias Mitigation Strategies

In order to address the second research question, various biases in the field of recommender system and bias mitigation techniques were discussed. In 1996, Friedman and Nissenbaum have categorized biases in computer systems into:

1. pre-existing bias, which represents every bias that existed before implementation,
2. technical bias, which arises from the way in which technical structures are inherently built, as well as
3. emergent bias, which arises during the process of using the computer system.

In order to discern which biases are the most relevant in the context of recommender systems another categorization has been made based on the summary provided by Ashokan and Haas in *Fairness metrics and bias mitigation strategies for rating predictions* (Ashokan & Haas, 2021). In this paper, biases were categorized into different pipelines of the recommender system building process: data generation, model building, as well as deployment and user interaction. The biases included in these pipelines were selected during the first iterations of viewing the corpus in order to reduce complexity, as the original list by the authors consisted of 21 bias types (Ashokan & Haas, 2021, p. 4). The selected biases have been discussed in the theory chapter, and they have a wide range of effects and causes. Lastly, bias mitigation types have been categorized into pre-, in-, and post-processing techniques.

4.2. Frequencies Assessed

The following nested list displays the frequencies of each category as well as their subcategories and specifications. Note that some of the categories allow the option of papers belonging to different subcategories and multiple specifications at once. Additionally, for purposes of transparency, a full table with all category values has been provided in the appendix. Graphs have been provided for each category in order to visualize the frequencies of each subcategory or attribute.

Fairness measure

Multiple categories allowed: yes

- 0. undefined: 0
- 1. individual fairness: 1
- 2. group fairness: 17
- 3. agnostic/flexible: 6
    - PA statistical parity: 5
    - OD equalized odds: 2
    - OP. equalized opportunity: 4
    - DI disparate impact: 4
    - CF counterfactual fairness: 3
    - CA. calibration fairness: 1
    - OT. other: 4
        - U: user-side: 18
        - P: provider-side: 5
            - *U+P: user- and provider-side:* 2
        - I: item-side: 2
            - singular. singular attributes only: 13
            - multiple. multiple attributes possible: 5

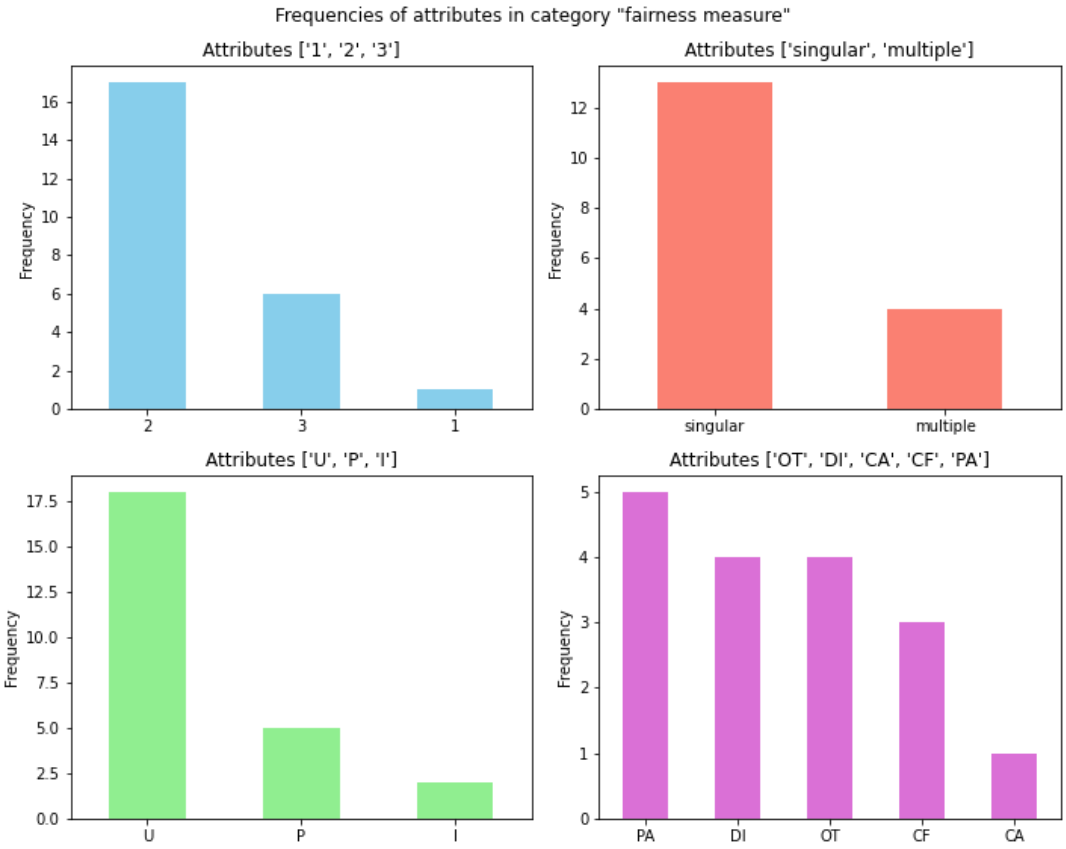

Fig. 1: Frequencies of attributes in category "fairness measure"

framing of source of discrimination

Multiple categories allowed: no

- 0. undefined: 11
- 1. lack of inclusion: 0
- 2. bad actors: 9
- 3. unconscious bias: 1
- 4. issue of distribution: 0
- 5. systemic issue: 2
- 6. superiority of computer systems over humans: 1

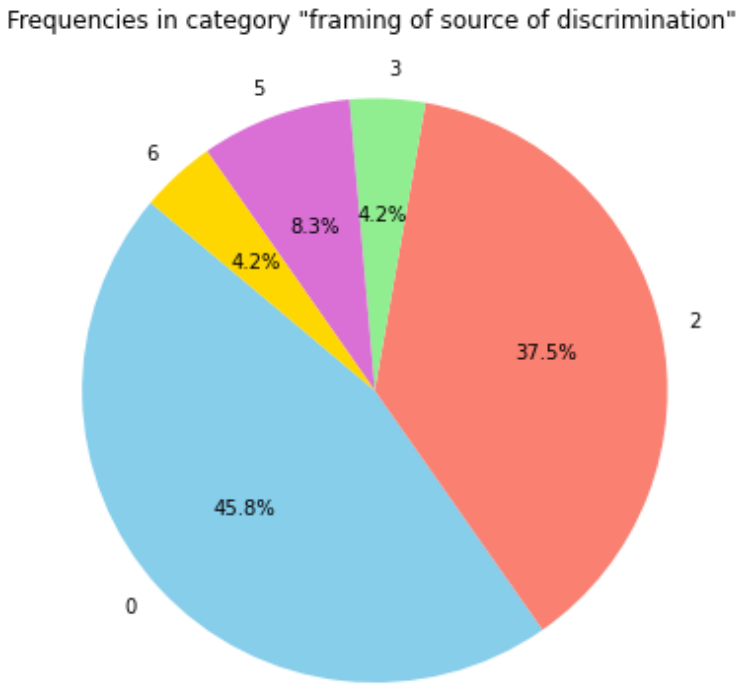

Fig. 2: Frequencies in category "framing of source of discrimination"

gender definition: used vs. proposed

Multiple categories allowed: no

- 0. undefined: 9
- 1.1. binary: 14
    - *proposed non-binary framework in the future: 1*
- 1.2 non-binary: 2

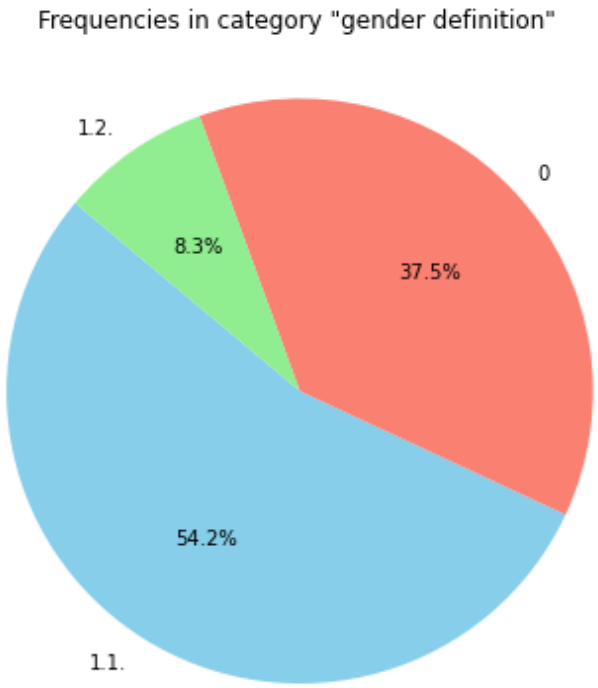

Fig. 3: Frequencies in category "gender definition"

type of bias treated

Multiple categories allowed: yes

- 0. undefined: 2
- data generation biases: 9
    - 1.1 historical bias: 4
    - 1.2. representational bias: 4
    - 1.3 simpson's paradox: 1
- model building biases - 13
    - 2.1. popularity bias: 4
    - 2.2. algorithmic bias: 3
    - 2.3. demographic bias: 6
- deployment and user interaction - 5
    - 3.1. social bias: 1
    - 3.2. interaction bias: 1
    - 3.3. ranking bias: 2
    - 3.4. filter bubbles: 1

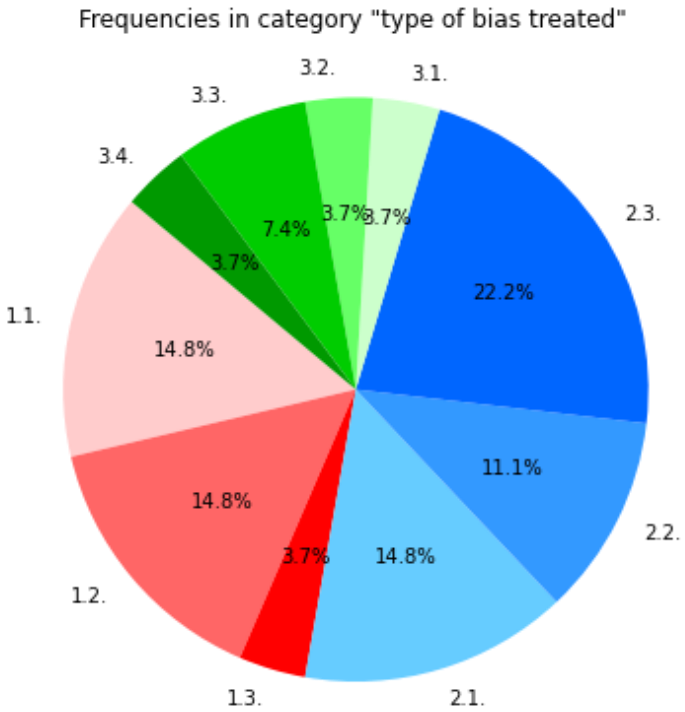

Fig. 4: Frequencies in category "type of bias treated"

bias mitigation method type

Multiple categories allowed: no

- 0. undefined: 1
- 1. pre-processing: 1
- 2. in-processing: 9
- 3. post-processing: 8
- 4. multiple: 5
    - *both in- and post-processing:* 3

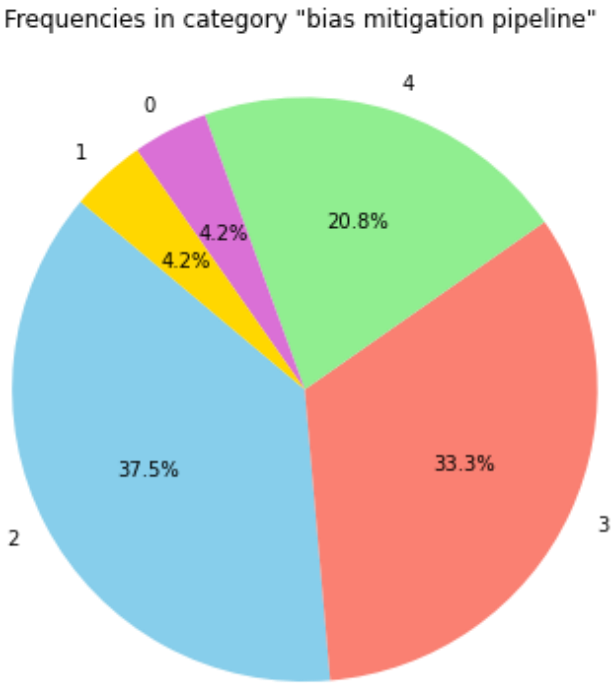

Fig 5: Frequencies in category "bias mitigation pipeline"

4.3. Secondary results

Fairness measures

The list of frequencies shows that in terms of stakeholders, the vast majority of research papers (75%) focus on user-side fairness. Item-side fairness was only included in 2 out of the 24 papers (8,33%). Provider fairness was addressed in 20,83% of the papers. Note that a paper may focus on fairness with respect to multiple stakeholders, which also happened in 2 cases (8,33%) and both papers were focused on user as well as provider fairness. For example, Wu et. al consider both users and providers as stakeholders when it comes to the fairness measure they used, extending on an already existing framework which was only focused on user fairness: "We extend their proposed framework to formalize a family of exposure fairness metrics that model

the problem jointly from the perspective of both the consumers and producers.“ (Wu et al., 2022, p. 703)

The papers whose fairness measures consisted of some sort of group split based on an attribute value were denoted with a specification regarding whether it is possible for one user, item or provider to belong to multiple protected attribute groups at once. 72,22% of the aforementioned research paper subgroup allowed only singular attributes.

When it comes to specific types of fairness metrics, statistical parity was the most popular with 20,83% of all research papers in the corpus including some variation of statistical parity in their fairness measure. However, equalized opportunity, disparate impact, and the category ‘other metrics’ appeared in 4 out of 24, or 16,67%, of cases. Overall, the metrics were distributed relatively evenly, with calibration fairness being the exception since it appeared only once in the corpus (4,17%).

Framing of source of discrimination

Out of the papers which had some sort of discernable framing of the origin of discrimination, bad actors (which includes individuals, data, and algorithms) were deemed at fault for discrimination in 69,23% of cases. Only 2 papers (15,38%) explicitly framed discrimination as a systemic issue to a sufficient degree. For example, Wu et. al have framed the integration of both user and provider fairness as an effort to mitigate a systemic issue of unfairness:

> “we argue in this work that joint consideration of group attributes on both user-side and item-side allows us to study other forms of systematic unfairness of social and moral import in recommendation. In this setting, the choice of group attributes on both sides is an important consideration and must be informed by historical and social contexts as well as critical scholarship in the area of socioeconomic justice.“ (Wu et al., 2022, p. 712)

Gender definition

37,5% of papers did not define gender sufficiently, that is as more than as an option for a variable.

Type of bias treated

The least attention was given to biases in the deployment and user interaction pipeline (20,83%). More biases were treated which belonged to the model building pipeline (54,16%) than the data generation pipeline (37,5%).

Bias mitigation method type

In- and post-processing techniques were used about the same amount of time, with in-processing being used 37,5% of the time and post-processing techniques being used a third of the time (33,33%). This excludes papers which use more than one bias mitigation method type at once (20,83%). 60% of papers who do use multiple bias mitigation types involve both in- and post-processing. For example, Ashokan and Haas propose an algorithm which "can be applied both as an in- processing approach by learning the adjustments during training procedure and applying them on the separate test data, or as post-processing approach which directly calculates the necessary adjustments on the test data itself" (Ashokan & Haas, 2021, p. 10).

4.4. Primary results

Fairness measures

17 out of 24 (70,83%) of the papers included a fairness measure based on a group split based on an attribute value such as gender, race, age, as well as other non-protected attributes (continent of origin) and placeholder values (majority vs. minority group, advantaged vs. disadvantaged group, etc.). Only 1 out of 24 papers was explicitly focused on individual fairness as a fairness measure. Lastly, a quarter of the papers use some sort of agnostic framework or consider multiple fairness measure types at once.

Framing of source of discrimination

Only 8,33% of research papers in the corpus explicitly define discrimination as a systemic issue. It is to be noted that almost half of the papers (45,83%) do not address the origin of discrimination to a sufficient degree in the first place.

Gender definition

Gender has been defined as a binary variable in 54,2% of cases where it was addressed to a sufficient degree. One of the papers using a binary gender variable proposed to include a non-binary gender structure in future work. Boratto et. al explained that their dataset was binary, but that their proposed framework could still be applied to a non-binary gender setting in the future:

> „Experiments were based on a binary gender construct, with datasets providing only two genders, ‚male' and ‚female'. Despite we had actually no chance of considering 'non-binary' constructs, our formulation can be still applied to attributes with more than two genders. We remind readers to (Hamidi et al. 2018) for consideration on the possible consequences of gender inference." (Boratto et al., 2021, p. 446)

Type of bias treated

The biases treated mostly belong to the model building pipeline category (54,17%). Out of those papers, 46,15% dealt with demographic bias as one of the main biases treated by their bias mitigation strategy.

Bias mitigation method type

Only 1 out of 24 papers (4,16%) included a pre-processing step as their main bias mitigation strategy.

5. Discussion

5.1. Interpretation and connection to current state of research

This section will provide the interpretative work based on the results of the qualitative content analysis for each of the chosen dimensions respectively. It will furthermore tie in the results of the study with the current state of research and provide thoughts on the fairness-accuracy trade-off.

5.1.1. Fairness Measures

The primary results show that the main focus of the research in recommender system fairness lies on group fairness measures. Also, more research papers included agnostic frameworks than individual fairness based fairness measures. This can be interpreted as researchers aiming to acknowledge unfairness with respect to protected groups, as opposed to the 'fairness through unawareness' approach which some instances of individual fairness fall under. This would connect with the current state of research as it represents the notion that fairness measures which bypass protected attributes allow for correlating attributes to manifest, resulting in unintentional discrimination (Apfelbaum et al., 2010, p. 907). It is to be noted that the one paper denoted as purely individual fairness based was categorized as '1.I', or 'individual item fairness'. This might be due to the idea that, intuitively, items to be recommended don't seem to have protected attributes in the traditional sense, i.e. a characteristic of a person which is protected under law. However, as Zhu et al. have noted, political ideology could be considered a protected attribute of an item such as the news which can be recommended to a user (Zhu et al., 2020, p. 452).

As the secondary results have shown, research in the field of recommender system fairness is mostly focused on user-side fairness. Provider fairness is only considered in 20,83% cases, and fairness on item-side was only featured twice in the corpus (8,33%). This might have many reasons. First of all, there might be systematic reasons for why fairness on user-side is deemed as a more urgent or serious problem than fairness on the side of providers on recommender platforms. More specifically, it might capture the neoliberal notion that providers on a platform are competing against each other, and therefore, not much intervention is needed in order to assert fairness amongst providers. The second reason as to why provider and item fairness are underrepresented in this field of research has been previously mentioned: A one-to-one mapping between an item and the individual who provides that item is not possible in most

cases, and an item might be mapped to more than one provider, which makes the representation of the providers' protected attribute(s) more difficult (Boratto et al., 2021, p. 426).

Furthermore, it remains to be discussed how the results of the study regarding fairness measures compare with the fairness definition derived by Hoffmann's *Where Fairness Fails.* The strong tendency towards using single-attribute-variables when it comes to group fairness measures indicates that the current state of research is far away from an intersectional approach, as has been advocated by Hoffmann, which would be ideally aware of the conditions which make privilege and advantage possible, as opposed to centering disadvantage. The latter seems to be especially apparent in the common usage of not using protected attributes directly, but using placeholder group titles such as 'disadvantaged' vs. 'advantaged' group instead. It is worth reiterating that even the use of multiple attributes would not fully account for what Hoffman would deem an 'intersectional approach', since "[i]ntersectionality is not a matter of randomly combining infinite variables to see what ‚disadvantages' fall out; rather, it is about mapping the production and contingency of social categories." (Hoffmann, 2019, p. 906)

As for the concrete group fairness metrics, statistical or demographic parity was the most common. Although not being easily compatible with the use of multiple attributes at once, and despite its tendency to put individual fairness at stake for the sake of group fairness, statistical parity remains as one of the most standard fairness metric. The same seems to be true for equalized odds and equalized opportunity, which are stricter versions of statistical parity. This might be due to the fact that equalized odds and opportunity are considered as relatively 'non-interventionalist' approaches (MIT, 2020). Why, how and when a measure of fairness ought to 'intervene' is a complex and context-specific question which shall be left to be discussed in future work, as it would exceed the frame of this master thesis. Furthermore, Castelnovo et al. have described that the issue with statistical parity is that it does not account for the reasoning behind *why* some distributions are shaped the way that they are, asking "if it is true that women repay their loans with higher probability, is it really fair to have demographic parity between men and women?" (Castelnovo et al., 2022, p. 7) Tracing and contextualizing bias seems to remain a difficult problem in recommender system fairness research.

Lastly, the reasoning behind why calibration fairness was only featured once (4,17%) in the corpus will be attributed to the fact that its main purpose is to combat one specific type of bias, which is popularity bias (Abdollahpouri et al., 2019).

5.1.2. Framing of Source of Discrimination

The primary results of this analysis show that for almost half of the research papers (45,83%), the topic of discrimination was not directly mentioned or addressed in depth, and only 8,33% defined the origins of discrimination as a systemic issue. These results reflect the tendency of leaving an in-depth discussion of the origin of biases in data, algorithms, and user interaction untouched, as has been discussed by Hoffmann. This can also bee seen as an instance of the phenomenon Sara Hooker has described in *Moving beyond 'algorithmic bias is a data problem'* of researchers in the field of computer science chronically diffusing the origin of bias, labeling it as 'somebody else's problem.'

The secondary results show that amongst those papers which did discuss the source of discrimination, most framed biased 'actors' at being at fault for the bias they treat. As previously discussed, Hoffmann criticizes this approach for a multitude of reasons: First, it frames discrimination in narrow cause-and-effect terms, blaming individual 'misguided' actors instead of positioning it in a societal context. Secondly, it therefore aims to cure individual actors (such as individuals, data, or algorithms) of its biased tendencies instead of addressing all contributing factors which have lead to this act of discrimination. In summary, this approach fails to admit to the structuring role of technology *within society.*

Interestingly, there were no papers which were categorized as framing the source of discrimination as due to 'lack of inclusion' or as an 'issue of distribution'. This might be due to the fact that, since almost half of the papers in the corpus did not address discrimination to a sufficient degree, the discussion of discrimination in these papers was shallow in general. Alternatively, this could be interpreted as having categories which were not as well-defined as previously anticipated. The 'get more data to be more inclusive' approach generally did not come up in a literal sense in the 16,66% of papers which aim to mitigate representational bias, i.e. bias caused by skewed datasets. This is reflected in the fact that those papers chose to mitigate representational bias mostly by a post-processing approach such as re-ranking, or via an in-processing strategy, or both. As for the 'issue of distribution' framing, it was not suggested by any of the papers that the redistribution of resources would be the solution against bias and discrimination. This might be due to the fact that this is a generally hard problem to address as well as solve on a technical level, in the field of computer science.

5.1.3. Gender Definition

The main results of the quantitative content analysis have shown that in the majority of cases (54,2%), gender is defined as a binary variable, where one of those papers proposed to include a broader gender definition in the future. The tendency towards declaring gender as a binary variable reflects a normative concept of gender which is performatively reconstructed and regulated with each paper in this category. To reiterate, this performative act is what creates "[g]ender reality" (Butler, 1990, p. 278) according to Butler, which is "real only to the extent that it is performed." (ibid.) Moreover, the lack of engagement with gender and homogenous usage of a binary gender definition hints less towards a perspective which Keyes et al. described, which would involve gender multiplicity, and more towards what they called a practice which treats gender variations as "second-order phenomena" (Keyes et al., 2021, p. 8). That being said, one of the papers which fell into the 'non-binary gender definition' category was *When Biased Humans Meet Debiased AI: A Case Study in College Major Recommendation*, which had a gender framing involving the options of 'male', 'female', 'non-binary', and an opt-out option for individuals who chose not to disclose their gender. It is to be noted that, although this was not included in the category system, this text was the only research paper involving a qualitative survey. This might be because quantitative research in computer science tends to be less likely to be guided by the principles of gender multiplicity. Research in recommender system fairness requires datasets, and in many cases, those datasets are standard public datasets. In some of the papers, such as *Causal Collaborative Filtering*, it has briefly come up that gender is defined as binary because that is how the dataset defines it. However, these cases exemplify exactly what Keyes et al. described as the construction of gender variation as a second-order phenomenon.

As the secondary results have shown, 37,5% of papers did not define gender sufficiently. This category seems to be positively correlated with not defining a 'source of discrimination', as more than half of those papers who did not define gender beyond a variable (55,56%) also did not go into a deeper discussion about discrimination, bias, and their origins. This would further tie into the notion that, overall, fairness remains to be a discursively shallow topic in the field of recommender system research. Three notions are important to reiterate and connect with these results: First, Hoffmann's term of discursive violence, which states that violence can be discursively discharged and result in creating the possibility of *othering*, such as individuals who do not fit into the gender binary. Second, Keyes et al.'s idea of ignorance being a social

phenomenon which is actively and continuously cultivated as opposed to merely a failure to generate knowledge about gender. And third, Pinney et al. warned about the pitfall of research dedicated towards system auditing which means well but still reproduce harm by perpetuating binary notions of gender, ultimately resulting in the system being audited for individuals who fall within the norm.

5.1.4. Type of Bias Treated

The primary results of the content analysis regarding the dimension of bias types has shown that recommender system fairness research mostly focuses on biases stemming from the model building process, mainly demographic bias. Moreover, the data shows that there exists a tendency towards solving problems caused by algorithms with strategies that affect it directly within the building or training process, as 53,85% of model building biases are dealt with via an in-processing technique. Demographic bias is mitigated via a group fairness measure 66,67% of the time. The focus on bias caused by the model itself suggests that ensuring the fairness of algorithms is (at least slightly) more of an interest in this field than ensuring fairness on a data-level. This is generally paired with the assumption that in the research paper's scenario, the data is already unbiased.

Furthermore, the secondary results show that there seems to be relatively little focus going into biases that appear after deployment and on the level of user interaction. This could be the result of post-deployment biases being delegated to a different area of expertise and therefore deemed as 'somebody else's problem', as has been described by Sara Hooker. When connecting these results with Hooker's insights, it appears as if bias in the field of recommender system fairness is not entirely viewed as a 'data problem', and that bias as an 'algorithm problem' is becoming more of a prevalent viewpoint.

5.1.5. Bias Mitigation Pipelines

Interestingly, the primary results of the qualitative content analysis shows that even in cases where there exists bias in the data, it is more likely for it to be mitigated via an in- or post-processing strategy than a pre-processing strategy. Similarly, the secondary results show that even in cases where there was more that one type of bias mitigation strategy, it probably a combination of in- and post-processing strategies. These results suggest that re-weighting and re-sampling strategies are not as common as described in Sara Hooker's article.

It is furthermore to be noted that the one paper which performed a pre-processing step in order to debias their recommender system was *When Biased Humans Meet Debiased AI: A Case Study in College Major Recommendation* by Chang et al.. In this study, gender bias was removed from user embeddings as a debiasing step, and then those embeddings were used to train the recommender model (Wang et al., 2023, p. 8). The survey results of Chang et al.'s study have shown that even though this pre-processing step preserved both user fairness through removing gender bias *and* the accuracy of the recommender system, the participants of the study preferred the biased recommender system. This ties into a complicated debate which will be briefly discussed in the last section of the discussion chapter titled the accuracy-fairness-tradeoff.

5.2. The Fairness-Accuracy Trade-Off

In her article, Sara Hooker mentioned various questions which have come up in recent years in the context of trade-offs on algorithmic bias, such as: "How does optimizing for compactness impact robustness and fairness? What about the trade-off between privacy and fairness?" (Hooker, 2021, p. 2) As a form of recommended action, the aim of this chapter is to build upon these questions and sensitize future researchers in the field of fair recommender systems to the implications of the fairness-accuracy trade-off. Traditionally, fairness in recommender systems is thought of as a *roadblock* to recommender accuracy. Many of the reviewed papers in this work, including *When Biased Humans Meet Debiased AI: A Case Study in College Major Recommendation,* are set on keeping the promise that their "debiased recommender makes fairer […] recommendations without sacrificing its accuracy in prediction." (Wang et al., 2023, p. 1) However, even in a case where both the biased and debiased system perform similarly in terms of accuracy, a follow-up survey showed that research participants preferred biased college major recommendations. Wang et al. conclude that "we cannot fully address the gender bias issue in AI recommendations without addressing the gender bias in humans." (ibid.)

First of all, these results pose the question whether the aim of fair recommender systems to increase fairness while remaining accurate in terms of prediction is productive when placed into a biased society. Secondly, framing fairness as a counterforce to accuracy inadvertently conveys the notion of fairness as an afterthought: 'Accuracy first, fairness second.' This seems to be especially true when contextualized within a competitive virtual environment where success is measured in terms of engagement. Given these insights, two movements are urgently needed: First, as Wang et al. have suggested, explainable AI might help with reconciling

debiased recommender systems with a biased user base, envisioning “an interactive and iterative human-AI bias co-training process where AI and humans work together iteratively and continuously to help correcting the biases in each other.“ (Wang et al., 2023, p. 24) Secondly, there ought to be a discussion around which impacts are prioritized when it comes to recommender system fairness: is it decreasing bias? Maintaining prediction accuracy? Gaining public endorsement? Increasing serendipity? Which of these aspects are treated as a given, which are treated as an afterthought, and why?

6. Conclusion

The first half of this thesis was dedicated towards introducing multiple theoretical concepts which can be attributed more broadly to field of Critical Algorithm and Data Studies. Multiple fairness definitions were introduced (individual fairness, statistical parity, equalized odds and equalized opportunity, disparate mistreatment and disparate impact, calibration fairness, and counterfactional fairness) as well as stakeholders which are affected by bias and discrimination on recommender platforms. In order to provide a context on a discursive level, several notions and pitfalls of fairness and antidiscrimination discourse in the field of computer science have been investigated through the work of Hoffmann and Keyes. This was followed by the introduction of several bias frameworks based on the work of Friedman and Nissenbaum, as well as machine learning pipelines. Gender definitions were provided with theoretical concepts by Judith Butler, Keyes et al, and Pinney et al. The last section of the theoretical foundation was dedicated towards explaining bias mitigation pipelines. This was followed by the methodological foundation, which explained the research process of Mayring's qualitative content analysis and how it was applied to the corpus of 24 samples in recommender system fairness literature. Based on the results of this study and the subsequent discussion, the next and final section of this thesis will be dedicated towards explicitly answering the previously defined research questions.

6.1. Summary of Answers to Research Questions

RQ1: How are the concepts of fairness and gender currently defined in recommender system research?

Fairness in this field seems to be currently defined mainly in terms of group fairness, whereas individual fairness by itself is not regarded as relevant. This means that fairness is generally being measured with respect to groups of certain protected attributes, either explicitly (groups which hold specific values of gender, age, race, or other attributes) or implicitly (groups which are categorized as 'advantaged' and 'disadvantaged'). When it comes to specific fairness measures, there seems to be no singular popular outlier, as the distribution of fairness measures was relatively even. A more specific fairness definition would therefore undermine the variety of measures used in this field. However, variations of statistical parity, equalized odds, and equalized opportunity were commonly seen, with statistical parity being slightly more common than the others. Potential issues with that have been discussed, as measures such as statistical

parity may not fully grasp the origin of bias and its contexts. Lastly, it was shown that in the vast majority of cases, fairness is seen as user-fairness specifically. Other stakeholders, such as providers or the items themselves, were rarely considered. This might be due to the difficulty of mapping providers to their respective protected attributes or due to the complexity of mapping attributes to multiple providers of a singular item (Boratto et al., 2021, p. 426). It might also be traced back to the simple concept that providers on a platform are simply not seen as important stakeholders when compared to users, which can be understood as representative of providers being historically underrepresented in fairness issues, as has been mentioned by Gómez et al (Gómez, Shui Zhang, et al., 2022, p. 435).

Gender was, in the majority of cases, defined explicitly as a binary variable, despite some efforts to acknowledge it as non-binary. Moreover, 37,5% of papers in the corpus did not define gender in general, that is: gender was defined as an option for a variable, but no further information was given. Only two papers operationalized the gender variable with more than two options. This falls into the narrative which Keyes et al. have described in *Much Ado About Gender*, which is that gender variations are commonly treated as "second-order phenomena" (Pinney et al., 2023, p. 8) in computer science. Furthermore, the inability for most fairness measures to account for multiple protected variables at once shows that recommender system research struggles to incorporate an intersectional approach.

RQ2: Which biases are most commonly treated and which bias mitigation strategies are currently the most common in the domain of recommender systems?

Recommender system research is currently mostly focused on biases originating in the model building pipeline, with biases most commonly falling under the category of demographic bias. Biases stemming from the data generation and user deployment pipelines are treated less in comparison. Furthermore, more than half of biases stemming from the model building process are also being treated with in-processing techniques, and fairness in this regard is most commonly measured with group fairness measures. Overall, in- and post-processing strategies are most commonly used in order to mitigate bias, with pre-processing strategies being rarely considered. The one example where pre-processing was indeed used was in order to de-bias word embeddings for a college major recommender (Wang et al., 2023).

When compared to the issues presented by Hoffmann regarding liberal antidiscrimination discourse in *Where Fairness Fails*, it seems as if discrimination seems indeed to be not mentioned in depth on a discursive level, and in those cases where a deeper discussion is discernable, it is mostly framed in terms of 'bad actors', that is to say: the issues of bias are due to algorithms behaving wrongly and datasets being skewed, as opposed to algorithms and datasets being embedded within societal power structures.

6.2. Limits

It is worth mentioning that there are multiple limits when it comes to the study conducted in this master thesis. First, it must be acknowledged that the corpus was small and contained only 24 papers. Although this was done in order to keep the papers limited to the subject matter (fairness and bias mitigation specifically in the area of recommender system research), a broader study concerning a wider range of fields might be in order, as was done in *Much Ado About Gender*, which was focused on Information Access Systems in general. The size of the corpus also dictates insights that can be retrieved via quantitative analysis when considering the Law of Large Numbers. Secondly, the gender dimension could have been analyzed in more detail, perhaps as detailed as Pinney et al. have done it. To reiterate, the authors recommended using an inclusive concept of gender in research which is based on a precise documentation of the gender variable collection process. Using gender variables in the context of auditing systems for fairness was stated as a positive way to include gender in computer science, as opposed using gender variables for the sake of gender prediction or gender personalization. A further investigation into what research goals were present in the corpus containing recommender system research literature would have been ideal. However, in order to reduce complexity in this study, the gender dimension was mainly analyzed on the basis of whether researchers operationalize gender as a binary or as a non-binary feature. Still, it is to be noted that there are many other aspects of gender definition in research which can be examined, as has been suggested by Keyes et al. in *You Keep Using That Word.* Third, it must be acknowledged that the category of 'framing of source of discrimination' was relatively shallow compared to the other categories, as it was difficult to discern when a paper is engaging in antidiscrimination discourse 'to a sufficient degree'. However, it was an attempt to follow Hoffmann's understanding of antidiscrimination discourse and its pitfalls. More research would be needed in the future in order to be able to quantify discursive structures in a more precise way. Lastly, as I am currently still in the midst of my Bachelor program for Artificial Intelligence, the framework of fairness metrics introduced in this paper was built upon a

beginner's level of understanding. A closer look at fairness measures and their politics would be needed, as has been suggested by Narayanan (Narayanan, 2019). Castelnovo et al. further provide a framework towards conceptualizing the nuances of fairness metrics in *A clarification of the nuances in the fairness metrics landscape.* A deeper understanding of the intricacies of this topic might reveal new and more nuanced insights in the field of recommender system fairness.

6.3. Prospect and Scientific Contribution

It remains to discuss the prospect regarding the development of fairness, bias and gender concepts in recommender system research. The results have shown that there is still much to be done in the area of Algorithm Fairness, including the development of robust and thoughtful fairness measures as well as discussions on the societal context in which bias occurs and the structuring role of technology in society. Two normative trends seem to be going strong: first, fairness is most commonly defined as user fairness, and providers (and their respective items, wherever that term is applicable) are not treated as valuable stakeholders in recommender system fairness. However, some papers have shown effort in explicitly including providers in the task of making recommender systems fair. Secondly, treating gender as a binary value seems to still be the most dominant practice. Nonetheless, in contrast to the study conducted by Pinney et al., there were two papers which operationalized gender as a variable more than two values. For example, one of it used the options of 'male', 'female', 'non-binary', and 'do not want to disclose' (Wang et al., 2023). Since this was done in the context of a qualitative survey, the question remains if this practice is easily transferable to quantitative methods. It is to be expected that, as Algorithm Fairness is going continues to grow as an important field of study in the intersection between computer science and social studies, these matters will be persistently treated with more urgency – especially as more and more cases arise in which recommender systems fail to act fairly.

The purpose of this work was to advance knowledge in the area of Algorithm Fairness by exploring the state of research in the field of fair recommender systems. Any effort within a thoughtful antidiscrimination discourse must first be based on a solid understanding of how terms such as fairness or gender are defined in the first place. Furthermore, the explication of which trends currently exist with respect to bias mitigation has aimed at directing attention towards which biases and which mitigation strategies might be currently overlooked. This work was built upon the massive efforts done by Hoffmann, Keyes et al., Pinney et al., Scheuerman

et al., and many more. It is their efforts, as as well as the aim of this thesis, to highlight the role which the production of knowledge plays within technological systems, as well as the importance of acknowledging modes of being which cannot be captured by datafication.

7. References

Abdollahpouri, H., Mansoury, M., Burke, R., & Mobasher, B. (2019). *The Unfairness of Popularity Bias in Recommendation*.

Akpinar, N.-J., DiCiccio, C., Nandy, P., & Basu, K. (2022). *Long-term Dynamics of Fairness Intervention in Connection Recommender Systems* Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, Oxford, United Kingdom. https://doi.org/10.1145/3514094.3534173

Apfelbaum, E. P., Pauker, K., Sommers, S. R., & Ambady, N. (2010). In Blind Pursuit of Racial Equality? *Psychol Sci*, *21*, 1592. https://doi.org/10.1177/0956797610384741

Ashokan, A., & Haas, C. (2021). Fairness metrics and bias mitigation strategies for rating predictions. *Inf. Process. Manage.*, *58*(5), 18. https://doi.org/10.1016/j.ipm.2021.102646

Azzopardi, L. (2021). *Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval* Proceedings of the 2021 Conference on Human Information Interaction and Retrieval, Canberra ACT, Australia. https://doi.org/10.1145/3406522.3446023

Baeza-Yates, R. (2018). Bias on the web. *Commun. ACM*, *61*(6), 54–61. https://doi.org/10.1145/3209581

Bagenstos, S. R. (2006). The Structural Turn and the Limits of Antidiscrimination Law. *California law review*, *94*, 47. https://doi.org/10.2307/20439026

Bauer, G. R., Braimoh, J., Scheim, A. I., & Dharma, C. (2017). Transgender-inclusive measures of sex/gender for population surveys: Mixed-methods evaluation and recommendations. *PLOS ONE*, *12*(5), 1-28. https://doi.org/10.1371/journal.pone.0178 (PLOS ONE)

Beer, D. (2009). Power through the algorithm? Participatory web cultures and the technological unconscious. *New media & society*, *11*(6), 985-1002. https://doi.org/10.1177/1461444809336551

Boratto, L., Carta, S., Iguider, W., Mulas, F., & Pilloni, P. (2022). Fair performance-based user recommendation in eCoaching systems. *User Modeling and User-Adapted Interaction*, *32*(5), 839–881. https://doi.org/10.1007/s11257-022-09339-6

Boratto, L., Fenu, G., & Marras, M. (2021). Interplay between upsampling and regularization for provider fairness in recommender systems. *User Modeling and User-Adapted Interaction*, *31*(3), 421–455. https://doi.org/10.1007/s11257-021-09294-8

Braman, S. (2009). *Change of state : information, policy, and power*. MIT Press.

Browne, S. (2015). *Dark Matters : On the Surveillance of Blackness*. Duke University Press. https://doi.org/10.1515/9780822375302
https://www.degruyter.com/isbn/9780822375302

Buolamwini, J., & Gebru, T. (2018). *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification* Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Proceedings of Machine Learning Research. https://proceedings.mlr.press/v81/buolamwini18a.html

Buonocore, T. (2019). Man is to Doctor as Woman is to Nurse: the Gender Bias of Word Embeddings. Why we should worry about gender inequality in Natural Language Processing techniques. *Towards Data Science*. https://towardsdatascience.com/gender-bias-word-embeddings-76d9806a0e17

Burke, R., Sonboli, N., & Ordonez-Gauger, A. (2018). *Balanced Neighborhoods for Multi-sided Fairness in Recommendation* Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Proceedings of Machine Learning Research. https://proceedings.mlr.press/v81/burke18a.html

Butler, J. (1990). Performative Acts and Gender Constitution: An Essay in Phenomenology and Feminist Theory. In S.-E. Case (Ed.), *Performing Feminisms: Feminist Critical Theory and Theatre*.

Butler, J. (1993). *Bodies that Matter: On the Discursive Limits of 'Sex'*. Routledge.

Cambridge, D. (2024). bias. In *Cambridge Dictionary*. Retrieved 23. 2. 24, from https://dictionary.cambridge.org/dictionary/english/bias

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., & Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, *12*(1), 4209. https://doi.org/10.1038/s41598-022-07939-1

Caton, S., & Haas, C. (2023). Fairness in Machine Learning: A Survey. *ACM Comput. Surv.* https://doi.org/10.1145/3616865

CFR, C. o. F. R. (1978). PART 1607 - UNIFORM GUIDELINES ON EMPLOYEE SELECTION PROCEDURES. In *Title 29 - Labor. Subtitle B - Regulations Relating to Labor (Continued). CHAPTER XIV - EQUAL EMPLOYMENT OPPORTUNITY COMMISSION.* (Vol. 4).

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*, *1989*(1, Article 8). http://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8

Damore, J. (2017). Google's Ideological Echo Chamber. How bias clouds our thinking about diversity and inclusion. In *WebArchive*. https://web.archive.org/web/20170809021151/https:/diversitymemo.com/

Dean, B. (2023, 28. 5.). We Analyzed 4 Million Google Search Results. Here's What We Learned About Organic Click Through Rate. *Backlinko*. https://backlinko.com/google-ctr-stats

Devinney, H., Björklund, J., & Björklund, H. (2022). *Theories of "Gender" in NLP Bias Research* Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea. https://doi.org/10.1145/3531146.3534627

DiCiccio, C., Hsu, B., Yu, Y., Nandy, P., & Basu, K. (2023). *Detection and Mitigation of Algorithmic Bias via Predictive Parity* Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, <conf-loc>, <city>Chicago</city>, <state>IL</state>, <country>USA</country>, </conf-loc>. https://doi.org/10.1145/3593013.3594117

Drozdowski, P., Rathgeb, C., Dantcheva, A., Damer, N., & Busch, C. (2020). *Demographic Bias in Biometrics:*

A Survey on an Emerging Challenge IEEE TRANSACTIONS ON TECHNOLOGY AND SOCIETY,

Ekstrand, M. D., Tian, M., Azpiazu, I. M., Ekstrand, J. D., Anuyah, O., McNeill, D., & Pera, M. S. (2018). *All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness* Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Proceedings of Machine Learning Research. https://proceedings.mlr.press/v81/ekstrand18b.html

Enloe, C. (2000). *Maneuvers : the international politics of militarizing women&apos:s lives*. University of California Press.

Epps-Darling, A., Bouyer, R. T., & Cramer, H. (2020). Artist Gender Representation in Music Streaming. Proceedings of the 21st International Society for Music Information Retrieval Conference. ISMIR,

Etherington, D. (2016). Baidu and KFC's new smart restaurant suggests what to order based on your face. *TechCrunch*. https://techcrunch.com/2016/12/23/baidu-and-kfcs-new-smart-restaurant-suggests-what-to-order-based-on-your-face/

Färber, M., Coutinho, M., & Yuan, S. (2023). Biases in scholarly recommender systems: impact, prevalence, and mitigation. *Scientometrics*, *128*(5), 2703–2736. https://doi.org/10.1007/s11192-023-04636-2

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). *Certifying and Removing Disparate Impact* Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia. https://doi.org/10.1145/2783258.2783311

Fleder, D., & Hosanagar, K. (2009). Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity. *Management Science*, *55*(5), 697-712. http://www.jstor.org/stable/40539182

Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks. In *Federal Probation* (Vol. 80, pp. 38–46).

Freeman, A. D. (1978). Legitimizing racial discrimination through antidiscrimination law: a critical review of Supreme Court doctrine. *Minnesota law review*, *62*, 1119.

Friedersdorf, C. (2017). The Most Common Error in Media Coverage of the Google Memo. *The Atlantic*. https://www.theatlantic.com/politics/archive/2017/08/the-most-common-error-in-coverage-of-the-google-memo/536181/

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, *14*, 347. https://doi.org/10.1145/230538.230561

Fu, Z., Xian, Y., Gao, R., Zhao, J., Huang, Q., Ge, Y., Xu, S., Geng, S., Shah, C., Zhang, Y., & Melo, G. d. (2020). *Fairness-Aware Explainable Recommendation over Knowledge Graphs* Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, China. https://doi.org/10.1145/3397271.3401051

Gershgorn, D. (2018). Amazons "holy grail" recruiting tool was actually just biased against women. . https://qz.com/1419228/amazons-ai-powered-recruiting-tool-was-biased-against-women.

Geyik, S. C., Ambler, S., & Kenthapadi, K. (2019). *Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search* Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA. https://doi.org/10.1145/3292500.3330691

Gómez, E., Boratto, L., & Salamó, M. (2022). Provider fairness across continents in collaborative recommender systems. *Inf. Process. Manage.*, *59*(1), 25. https://doi.org/10.1016/j.ipm.2021.102719

Gómez, E., Shui Zhang, C., Boratto, L., Salamó, M., & Ramos, G. (2022). Enabling cross-continent provider fairness in educational recommender systems. *Future Generation Computer Systems*, *127*, 435-447. https://doi.org/https://doi.org/10.1016/j.future.2021.08.025

Gomez-Uribe, C. A., & Hunt, N. (2016). The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Trans. Manage. Inf. Syst.*, *6*(4), Article 13. https://doi.org/10.1145/2843948

Gössl, S. (2023). Recommender Systems and Discrimination. In (pp. 13-29). https://doi.org/10.1007/978-3-031-34804-4_2

Gotanda, N. (1991). A Critique of "Our Constitution Is Color-Blind". *Stanford Law Review*, *44*(1), 1-68. https://doi.org/10.2307/1228940

Griggs v. Duke Power Co., 401 U.S. 424, (1971). https://supreme.justia.com/cases/federal/us/401/424/

Guo, H., Li, J., Wang, J., Liu, X., Wang, D., Hu, Z., Zhang, R., & Xue, H. (2023). *FairRec: Fairness Testing for Deep Recommender Systems* Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis, <conf-loc>, <city>Seattle</city>, <state>WA</state>, <country>USA</country>, </conf-loc>. https://doi.org/10.1145/3597926.3598058

Hamidi, F., Scheuerman, M. K., & Branham, S. M. (2018). *Gender Recognition or Gender Reductionism? The Social Implications of Embedded Gender Recognition Systems* Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, <conf-loc>, <city>Montreal QC</city>, <country>Canada</country>, </conf-loc>. https://doi.org/10.1145/3173574.3173582

Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning.

Hoffmann, A. L. (2019). Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, communication & society*, *22*, 915. https://doi.org/10.1080/1369118X.2019.1573912

Hoffmann, A. L. (2021). Terms of inclusion: Data, discourse, violence. *New media & society*, *23*, 3556. https://doi.org/10.1177/1461444820958725

Hooker, S. (2021). Moving beyond "algorithmic bias is a data problem". *Patterns*, *2*(4), 100241. https://doi.org/https://doi.org/10.1016/j.patter.2021.100241

Hoyningen-Huene, P. (1987). Context of discovery and context of justification. *Studies in History and Philosophy of Science Part A*, *18*(4), 501-515. https://doi.org/https://doi.org/10.1016/0039-3681(87)90005-7

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2017). Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. https://doi.org/10.48550/arxiv.1711.05144

Keyes, O. (2018). The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proc. ACM Hum.-Comput. Interact.*, *2*(CSCW), Article 88. https://doi.org/10.1145/3274357

Keyes, O. (2019, April 08, 2019). Counting the Countless. Why data science is a profound threat for queer people. *Real Life Magazine*. https://reallifemag.com/counting-the-countless/

Keyes, O., May, C., & Carrell, A. (2021). You Keep Using That Word: Ways of Thinking about Gender in Computing Research. *Proc. ACM Hum.-Comput. Interact.*, *5*(CSCW1), Article 39. https://doi.org/10.1145/3449113

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). *Avoiding discrimination through causal reasoning* Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA.

Kusner, M., Loftus, J., Russell, C., & Silva, R. (2017). *Counterfactual fairness* Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA.

Laws, M. (2020). Why we capitalize 'Black' (and not 'white'). *Columbia Journal Review*. https://www.cjr.org/analysis/capital-b-black-styleguide.php

LeCun, Y. [@ylecun]. (2020, 21. 6.). *ML systems are biased when data is biased. This face upsampling system makes everyone look white because the network was* [social media post]. X. Retrieved 23. 2. 2024 from https://twitter.com/ylecun/status/1274782757907030016?s=20

Liu, J. (2023). *Toward A Two-Sided Fairness Framework in Search and Recommendation* Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, Austin, TX, USA. https://doi.org/10.1145/3576840.3578332

Liu, Z., Fang, Y., & Wu, M. (2023). Mitigating Popularity Bias for Users and Items with Fairness-centric Adaptive Recommendation. *ACM Trans. Inf. Syst.*, *41*(3), Article 55. https://doi.org/10.1145/3564286

Masri, W., & Assi, R. A. (2014). Prevalence of coincidental correctness and mitigation of its impact on fault localization. *ACM Trans. Softw. Eng. Methodol.*, *23*(1), Article 8. https://doi.org/10.1145/2559932

Mayring, P. (2014). *Qualitative content analysis: theoretical foundation, basic procedures and software solution* [Monographie]. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-395173

Mazumder, P., & Singh, P. (2022). Protected attribute guided representation learning for bias mitigation in limited data. *Know.-Based Syst.*, *244*(C), 13. https://doi.org/10.1016/j.knosys.2022.108449

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, *54*(6), Article 115. https://doi.org/10.1145/3457607

MIT. (2020). *Fairness Criteria*. MIT. Retrieved 22. 2. from https://ocw.mit.edu/courses/res-ec-001-exploring-fairness-in-machine-learning-for-international-development-spring-2020/pages/module-three-framework/fairness-criteria/

Murphy, G. L. (2002). *The big book of concepts*. MIT Press.

Musto, C., Lops, P., & Semeraro, G. (2021). Fairness and Popularity Bias in Recommender Systems: an Empirical Evaluation. DP@AI*IA,

Nandy, P., DiCiccio, C., Venugopalan, D., Logan, H., Basu, K., & Karoui, N. E. (2022). *Achieving Fairness via Post-Processing in Web-Scale Recommender Systems* ✱ Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea. https://doi.org/10.1145/3531146.3533136

Narayanan, A. [A. Narayanan]. (2019). *Tutorial: 21 fairness definitions and their politics* [YouTube Video]. YouTube. Retrieved 29. 2. 24 from https://www.youtube.com/watch?v=jIXIuYdnyyk

Nguyen, T. T., Hui, P.-M., Harper, F. M., Terveen, L., & Konstan, J. A. (2014). *Exploring the filter bubble: the effect of using recommender systems on content diversity* Proceedings of the 23rd international conference on World wide web, Seoul, Korea. https://doi.org/10.1145/2566486.2568012

Nissenbaum, H. (2001). How Computer Systems Embody Values. *34*(03), 120,118-119. https://doi.org/10.1109/2.910905

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press. https://doi.org/10.2307/j.ctt1pwt9w5

Nyarko, J., Goel, S., & Sommers, R. (2021). *Breaking Taboos in Fair Machine Learning: An Experimental Study* Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, --, NY, USA. https://doi.org/10.1145/3465416.3483291

Pariser, E. (2011a). *Author Q&A with Eli Pariser*. ACM Ditgital Library. Retrieved 23. 3. from https://dl.acm.org/doi/book/10.5555/2029079

Pariser, E. (2011b). *The Filter Bubble: What The Internet Is Hiding From You*. Penguin Books Limited. https://books.google.at/books?id=-FWO0puw3nYC

Pinney, C., Raj, A., Hanna, A., & Ekstrand, M. D. (2023). *Much Ado About Gender: Current Practices and Future Recommendations for Appropriate Gender-Aware Information Access* Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, Austin, TX, USA. https://doi.org/10.1145/3576840.3578316

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On Fairness and Calibration. *Advances in Neural Information Processing Systems.*

Rahmani, H. A., Naghiaei, M., Dehghan, M., & Aliannejadi, M. (2022). *Experiments on Generalizability of User-Oriented Fairness in Recommender Systems* Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, <conf-loc>, <city>Madrid</city>, <country>Spain</country>, </conf-loc>. https://doi.org/10.1145/3477495.3531718

Ricci, F., Rokach, L., & Shapira, B. (2022). Recommender Systems: Techniques, Applications, and Challenges. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook* (pp. 1-35). Springer US. https://doi.org/10.1007/978-1-0716-2197-4_1

Scheuerman, M. K., Paul, J. M., & Brubaker, J. R. (2019). How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW), Article 144. https://doi.org/10.1145/3359246

Scheuerman, M. K., Spiel, K., Haimson, O. L., Hamidi, F., & Branham, S. M. (2020). *HCI Gender Guidelines*. Retrieved 18. 2. from https://www.morgan-klaus.com/gender-guidelines.html

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017). No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World.

Shelton, K. (2017, 30. 10.). The Value Of Search Result Rankings. *Forbes*. https://www.forbes.com/sites/forbesagencycouncil/2017/10/30/the-value-of-search-results-rankings/

SMC, T. S. M. C. *Critical Algorithm Studies: a Reading List*. Retrieved 19. 3. from https://socialmediacollective.org/reading-lists/critical-algorithm-studies/#1.1

Spade, D. (2015). *Normal life : administrative violence, critical trans politics, and the limits of law* (Revised and expanded edition. ed.). Duke University Press.

Star, S., & Bowker, G. (2007). Enacting silence: Residual categories as a challenge for ethics, information systems, and communication. *Ethics and Information Technology*, *9*, 273-280. https://doi.org/10.1007/s10676-007-9141-7

Statt, N. (2018). Google personalizes search results even when you're logged out, new study claims / A study, albeit from competitor DuckDuckGo, finds that Google search results can vary significantly. *The Verge*. https://www.theverge.com/2018/12/4/18124718/google-search-results-personalized-unique-duckduckgo-filter-bubble

Suresh, H., & Guttag, J. (2021). *A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle* Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, --, NY, USA. https://doi.org/10.1145/3465416.3483305

Verma, S., & Rubin, J. (2018, 29-29 May 2018). Fairness Definitions Explained. 2018 IEEE/ACM International Workshop on Software Fairness (FairWare),

Wadsworth, C., Vera, F., & Piech, C. (2018). *Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction* Arxiv. https://arxiv.org/abs/1807.00199

Wang, C., Wang, K., Bian, A. Y., Islam, R., Keya, K. N., Foulds, J., & Pan, S. (2023). When Biased Humans Meet Debiased AI: A Case Study in College Major Recommendation. *ACM Trans. Interact. Intell. Syst.*, *13*(3), Article 17. https://doi.org/10.1145/3611313

Wang, T., & Wang, D. (2014). Why Amazon's Ratings Might Mislead You: The Story of Herding Effects. *Big Data*, *2*(4), 196-204. https://doi.org/10.1089/big.2014.0063

Wang, W., Feng, F., Nie, L., & Chua, T.-S. (2022). *User-controllable Recommendation Against Filter Bubbles* Proceedings of the 45th International ACM SIGIR Conference

on Research and Development in Information Retrieval, <conf-loc>, <city>Madrid</city>, <country>Spain</country>, </conf-loc>. https://doi.org/10.1145/3477495.3532075

Wang, X., & Wang, W. H. (2022). *Providing Item-side Individual Fairness for Deep Recommender Systems* Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea. https://doi.org/10.1145/3531146.3533079

Watson v. Fort Worth Bank & Trust, 487 U.S. 977, (1988). https://supreme.justia.com/cases/federal/us/487/977/

Wei, T., & He, J. (2022). *Comprehensive Fair Meta-learned Recommender System* Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington DC, USA. https://doi.org/10.1145/3534678.3539269

Wu, H., Mitra, B., Ma, C., Diaz, F., & Liu, X. (2022). *Joint Multisided Exposure Fairness for Recommendation* Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, <conf-loc>, <city>Madrid</city>, <country>Spain</country>, </conf-loc>. https://doi.org/10.1145/3477495.3532007

Xu, S., Ge, Y., Li, Y., Fu, Z., Chen, X., & Zhang, Y. (2023). *Causal Collaborative Filtering* Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, <conf-loc>, <city>Taipei</city>, <country>Taiwan</country>, </conf-loc>. https://doi.org/10.1145/3578337.3605122

Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). *Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment* Proceedings of the 26th International Conference on World Wide Web, Perth, Australia. https://doi.org/10.1145/3038912.3052660

Zhu, Z., Wang, J., & Caverlee, J. (2020). *Measuring and Mitigating Item Under-Recommendation Bias in Personalized Ranking Systems* Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, China. https://doi.org/10.1145/3397271.3401177

8. Appendix

Table of Corpus:

index 1	venue	year	country of origin	publisher country
1	Information Processing and Management: an International Journal	2021	USA	USA
2	Information Processing and Management: an International Journal	2022	Spain, Italy	USA
3	FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency	2022	USA	USA
4	FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency	2023	USA, Switzerland	USA
5	ACM Transactions on Information Systems	2023	Singapore	USA
6	SIGIR '22: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval	2022	USA, UK, Netherlands, Iran	USA
7	AIES '22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society	2022	USA	USA
8	ISSTA 2023: Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis	2023	China	USA
9	Scientometrics	2023	Germany	Germany
10	User Modeling and User-Adapted Interaction	2021	Spain, Italy, Switzerland	USA
11	FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency	2022	USA	USA
12	Future Generation Computer Systems	2022	Spain, Italy, Portugal	Netherlands
13	KDD '22: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining	2022	USA	USA
14	SIGIR '20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval	2020	USA	USA
15	SIGIR '22: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval	2022	Canada, Hong Kong	USA
16	CHIIR '23: Proceedings of the 2023 Conference on Human Information Interaction and Retrieval	2023	USA	USA
17	KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining	2019	USA	USA
18	ACM Transactions on Interactive Intelligent Systems	2023	USA	USA
19	RecSys '22: Proceedings of the 16th ACM Conference on Recommender Systems	2022	Austria	USA
20	SIGIR '22: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval	2022	Singapore, China	USA
21	User Modeling and User-Adapted Interaction	2022	Italy	USA
22	Information Processing and Management: an International Journal	2017	Germany, Italy, Spain	USA
23	ICTIR '23: Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval	2023	USA, China	USA
24	SIGIR '20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval	2020	USA	USA

index 2	authors	title	doi
1	Ashwathy Ashokan, Christian Haas	Fairness metrics and bias mitigation strategies for rating predictions	https://doi.org/10.1016/j.ipm.2021.102646
2	Elizabeth Gómez, Ludovico Boratto, Maria Salamó	Provider fairness across continents in collaborative recommender systems	https://doi.org/10.1016/j.ipm.2021.102719
3	Xiuling Wang, Wendy Hui Wang	Providing Item-side Individual Fairness for Deep Recommender Systems	https://doi.org/10.1145/3531146.3533079
4	Cyrus DiCiccio, Brian Hsu, Yinyin Yu, Preetam Nandy, Kinjal Basu	Detection and Mitigation of Algorithmic Bias via Predictive Parity	https://doi.org/10.1145/3593013.3594117
5	Zhongzhou Liu, Yuan Fang, Min Wu	Mitigating Popularity Bias for Users and Items with Fairness-centric Adaptive Recommendation	https://doi.org/10.1145/3564286
6	Hossein A. Rahmani, Mohammadmehdi Naghiaei, Mahdi Dehghan, Mohammad Aliannejadi	Experiments on Generalizability of User-Oriented Fairness in Recommender Systems	https://doi.org/10.1145/3477495.3531718
7	Nil-Jana Akpinar, Cyrus DiCiccio, Preetam Nandy, Kinjal Basu	Long-term Dynamics of Fairness Intervention in Connection Recommender Systems	https://doi.org/10.1145/3514094.3534173
8	Huizhong Guo, Jinfeng Li, Jingyi Wang, Xiangyu Liu, Dongxia Wang, Zehong Hu, Rong Zhang, Hui Xue	FairRec: Fairness Testing for Deep Recommender Systems	https://doi.org/10.1145/3597926.3598058
9	Michael Färber, Melissa Coutinho, Shuzhou Yuan	Biases in scholarly recommender systems: impact, prevalence, and mitigation	https://doi.org/10.1007/s11192-023-04636-2
10	Ludovico Boratto, Gianni Fenu, Mirko Marras	Interplay between upsampling and regularization for provider fairness in recommender systems	https://doi.org/10.1007/s11257-021-09294-8
11	Preetam Nandy, Cyrus DiCiccio, Divya Venugopalan, Heloise Logan, Kinjal Basu, Noureddine El Karoui	Achieving Fairness via Post-Processing in Web-Scale Recommender Systems	https://doi.org/10.1145/3531146.3533136
12	Elizabeth Gómez, Carlos Shui Zhang, Ludovico Boratto, Maria Salamó, Guilherme Ramos	Enabling cross-continent provider fairness in educational recommender systems	https://doi.org/10.1016/j.future.2021.08.025
13	Tianxin Wei, Jingrui He	Comprehensive Fair Meta-learned Recommender System	https://doi.org/10.1145/3534678.3539269
14	Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, Gerard de Melo	Fairness-Aware Explainable Recommendation over Knowledge Graphs	https://dl.acm.org/doi/10.1145/3397271.3401051
15	Haolun Wu, Bhaskar Mitra, Chen Ma, Fernando Diaz, Xue Liu	Joint Multisided Exposure Fairness for Recommendation	https://dl.acm.org/doi/10.1145/3477495.3532007
16	Jiqun Liu	Toward A Two-Sided Fairness Framework in Search and Recommendation	https://dl.acm.org/doi/10.1145/3576840.3578332
17	Sahin Cem Geyik, Stuart Ambler, Krishnaram Kenthapadi	Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search	https://dl.acm.org/doi/10.1145/3292500.3330691
18	Clarice Wang, Kathryn Wang, Andrew Y. Bian, Rashidul Islam, Kamrun Naher Keya, James Foulds, Shimei Pan	When Biased Humans Meet Debiased AI: A Case Study in College Major Recommendation	https://dl.acm.org/doi/10.1145/3611313
19	Alessandro B. Melchiorre, Navid Rekabsaz, Christian Ganhör, Markus Schedl	ProtoMF: Prototype-based Matrix Factorization for Effective and Explainable Recommendations	https://dl.acm.org/doi/10.1145/3523227.3546756
20	Wenjie Wang, Fuli Feng, Liqiang Nie, Tat-Seng Chua	User-controllable Recommendation Against Filter Bubbles	https://dl.acm.org/doi/10.1145/3477495.3532075
21	Ludovico Boratto, Salvatore Carta, Walid Iguider, Fabrizio Mulas, Paolo Pilloni	Fair performance-based user recommendation in eCoaching systems	https://dl.acm.org/doi/10.1007/s11257-022-09339-6
22	Meike Zehlike, Tom Sühr, Ricardo Baeza-Yates, Francesco Bonchi, Carlos Castillo, Sara Hajian	Fair Top-k Ranking with multiple protected groups	https://doi.org/10.1016/j.ipm.2021.102707
23	Shuyuan Xu, Yingqiang Ge, Yunqi Li, Zuohui Fu, Xu Chen, Yongfeng Zhang	Causal Collaborative Filtering	https://dl.acm.org/doi/10.1145/3578337.3605122
24	Ziwei Zhu, Jianling Wang, James Caverlee	Measuring and Mitigating Item Under-Recommendation Bias in Personalized Ranking Systems	https://dl.acm.org/doi/10.1145/3397271.3401177

index 3	fairness measure	framing of source of discrimination	gender definition used	type of bias treated	bias mitigation method type
1	3.U.	2	1.1.	1.2.	4 (2, 3)
2	2.DI.P..singular	2	0	1.1.;1.2.	3
3	1.I.	0	0	2.1.	4 (2, 3)
4	2.PA.U.singular	0	0	2.2.	3
5	2.OT.	2	1.1.	2.1.	2
6	2.CA.U+P.singular	2	0	2.1.	3
7	2.PA.U.singular	0	1.1.	2.1.	2
8	2.OT.U.multiple	0	1.1.	0	4 (2, 3)
9	3.U.	2	1.1.	0	4
10	2.DI.P.singular	3	1.1. (proposed: 1.2.)	2.3.	2
11	2.OD+OP.U.singular	0	0	3.3.	3
12	2.DI.P.multiple	0	0	2.3.	3
13	2.CF.U.multiple	0	1.1.	2.3.	2
14	3.U.	0	0	1.2.	3
15	3.U+P.	5	0	1.1.	0
16	3.OD+OP.U.singular	6	0	1.1.;2.3.;3.2.	4
17	2.PA+OP.U.multiple	2	1.1. (proposed: 1.2.)	2.2.	2
18	3.U.singular	5	1.2.	1.1.; 2.3.;3.1.	1
19	2.OT.U.singular	0	1.1.	2.3.	2
20	2.CF.U.singular	2	1.1.	3.4.	2
21	2.PA+DI.U.singular	0	1.1.	3.3.	3
22	2.U.OT.multiple	2	1.1.	2.2.	3
23	2.CF.U.singular	0	1.1.	1.3.	2
24	2.PA+OP.I.singular	2	1.1.	1.2.	2