

Algorithmic Bias: An Integrative Review and Scope for Future Research

AMIT KUMAR CHAUDHARY

f20amitc@iimidr.ac.in


Indian Institute of Management Indore

Research Article

Keywords: Algorithmic bias, Artificial Intelligence, Machine Learning, Ethics, Consumers, Society. Organization

Posted Date: August 21st, 2024

DOI: <https://doi.org/10.21203/rs.3.rs-4775268/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Today Artificial Intelligence and Machine Learning (ML) algorithms are influencing various aspects of human life, for instance - healthcare, loan provision, education, recruitment, and so on. But these systems are facing the issue of algorithmic bias, they can potentially generate socially biased outcomes, and they can enhance inequalities in the workplace as well as in society, even when there is no intention of doing so. The current literature on algorithmic bias is progressing in various directions in the absence of a robust theoretical foundation. Therefore, there is a requirement for a consolidation to provide a comprehensive and up-to-date summary of research in the area. This study presents an integrative review of the current body of literature on algorithmic bias, considering the diverse domains, samples, and methodologies employed in previous studies. This analysis highlights multiple gaps in the algorithmic bias domain. These gaps comprise definitional issues, insufficient theoretical foundations, thematic tensions, and inconsistencies in current literature. A potential future research avenue is proposed, which consists of a collection of various themes and research gaps. Also, a theoretical framework is provided that might serve as a guiding principle for future research in the domain of algorithmic bias.

1. Introduction

Today, various organizations are adopting emerging technologies such as data analytics, big data, and Artificial Intelligence (AI) to transform and improve their decision-making and key operations (Kordzadeh & Ghasemaghaei, 2022). Specifically, organizations employ algorithmic systems to assist human decision-makers or automate decision-making processes, such as recommendation, predictive analysis, or classification. Machine Learning (ML) algorithms touch all aspects of human lives today. For instance, algorithms suggest recommendations for movies, products to be purchased, and people for dating. ML algorithms are also being used in various critical places such as policing, criminal justice, hiring decisions, loans, and so on. Algorithmic decision-making has many benefits, for instance, machines do not get tired or bored and are also capable of considering more factors than humans. However, similar to humans, algorithms are susceptible to biases which could make their decisions unfair. From a decision-making context, fairness means the absence of favouritism or bias towards a specific group or individual based on inherently acquired characteristics. Therefore, an algorithm is unfair when its decisions are skewed for specific groups or individuals.

One crucial example is software named COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) used for deciding on, pretrial detention or release of a person by judges in U.S. court. COMPAS predicts the risk of an individual committing a crime again. The software was claimed to be biased against African-American offenders, as it predicts them to have a higher risk of recommitting the crime than Caucasian offenders. Similarly, there are various other findings such as AI systems judging beauty pageants found to be biased against dark-skin contestants, and facial recognition systems more often predicting Asians to be blinking. These biases can enter in a system due to the use of imbalanced data or biased algorithms (Mehrabi et al., 2021).

Schwartz et al. (2022), discussed about human-centered design of AI, as it will improve users' capability to use AI efficiently and effectively. Human-in-loop approach is required to implement in order to deal with AI bias. For understanding the human-in-loop approach it is important to consider socio-technical factors, which include various topics such as human-AI interaction, psychology, organization behaviour, human factors, and so on. There is a need to develop guidelines for implementing a human-in-loop approach to reduce human and computational biases in a complex setting.

There is the prominence of topic algorithmic bias in both academics as well as mainstream discussions which encompasses various disciplines such as computer science, social science, legal and political studies, and humanities, having a lack of coherence. Hence there is a need to integrate these diverse perspectives and mechanisms through which algorithmic bias can be created, consumed, and disseminated. To obtain a comprehensive understanding of the algorithmic bias phenomenon, it is essential to synthesize the existing literature on the topic. To address this gap, this study is adopting an integrative literature review on the algorithmic bias topic.

An integrative literature review methodology involves the review, critical examination, and synthesis of representative literature on a topic (Torraco, 2005); integrative review is particularly suitable in cases where the existing research on a certain subject is fragmented across several domains and has not been systematically integrated and analyzed by scholars (Scully-Russ & Torraco, 2020). This is also the case with an existing body of algorithmic bias literature. In recent years, subsequent to the rise of the algorithmic bias phenomenon, few literature reviews have been conducted by scholars focusing on the existing state of algorithmic bias research. These reviews have exhibited a significant constraint by restricting themselves to either survey of preliminary research on topic (Arora, 2020; Kordzadeh & Ghasemaghahi, 2022), analysis of empirical research on fairness perception (Starke, Baleis, Keller & Marcinkowski, 2022), review focus on types and various definition of fair algorithm (Mehrabian et al., 2021) and detection techniques (Mehrabian et al., 2021; Drozdowski et al., 2020) or implication of algorithmic bias in specific domain such as education (Baker & Hawn, 2021), and healthcare (Arora, 2020).

This study asserts that a narrow approach has been taken by the aforementioned research and hence is unable to provide a comprehensive overview of past literature on algorithmic bias. Further, even though the literature on the topic of algorithmic bias has grown in recent years, empirical studies on the topic are scarce (Kordzadeh & Ghasemaghahi, 2022). Also, existing literature lacks a robust theoretical framework for providing guidance to research for further advancements in this field.

The objective of this study is to address the limitations present in prior literature reviews by using an integrative approach, hence making a valuable contribution to existing literature and advancing research in this domain. This study takes a broader view of algorithmic bias and encompasses literature not restricted to any specific fields, hence able to provide an up-to-date and comprehensive view of algorithmic bias topic. By engaging in this endeavour, the study presents a robust research profile and highlights the predominant themes explored in existing literature pertaining to the phenomena. A lack of thorough understanding of existing research on algorithmic bias may discourage scholars from working

in the domain of algorithmic bias, and practitioners will not be able to incorporate the acquired knowledge in order to deal with the issue of algorithmic bias. Therefore, in accordance with the objective of this research, the following are the study's research questions (RQs):

RQ1: What is the current status of Algorithmic bias literature?

RQ2: What are the existing research gaps, limitations, and suggestions for academics and practitioners?

With the objective of answering these research questions, this integrative review of algorithmic bias targets to comprehensively review all theoretical and empirical literature on algorithmic bias.

This examination included studies that have been published on the subject within the previous two decades, that is from 2001 onwards. To obtain scholars' contributions in the literature on algorithmic bias, the literature search was conducted across prominent databases for peer-reviewed literature. Although the term algorithmic bias was born in 1980 (Schwartz, 2019), studies in this domain began around 2011, therefore, to ensure that the literature review considers the publication lag, if any, with the aim to cover all the studies after 2001. This study is answering the RQ1, i.e., it identified, synthesized, and presented the current profile of algorithmic bias literature, it included coverage of studies, framework, variables, and measures pertaining to characterization and spreading of algorithmic bias. Subsequently, it addresses RQ 2, i.e., to explain the accumulated body of scholarly work about algorithmic bias from the beginning to the present, by doing so, this study presents detailed findings from selected published studies on algorithmic bias in peer-reviewed journals and highlights the limitations and gaps in current literature. Subsequently, this study offers potential direction for future research, and a framework for algorithmic bias research developed integrative review of existing scholarly work.

The remaining part of this paper is structured as follows: Section 2 explains the methodological process used to conduct this integrative review. Section 3 lists the important aspects of algorithmic bias literature. Section 4 discusses the aspects of the empirical literature on algorithmic fairness. Section 5 provides a research framework and recommendations for future researchers to address extant knowledge gaps. In section 6, the synthesis of limitations and gaps of existing studies are included, hence providing the themes for future research. Finally, the paper discussed the theoretical and practical implications of the study and also mentioned the limitations of the study.

2. Methodology

An integrated review summarizes existing empirical or theoretical research and offers a more thorough understanding of a specific phenomenon (Broome 1993). As a result, integrative reviews possess the ability to advance practice, policy, and research initiatives in specific domains. A properly conducted Integrative review provides the status of the research, contributes to theory, and has direct application to policy and practice (Whittemore & Knafl, 2005). Integrative reviews provide a significant opportunity to capture the existing knowledge on a certain topic to date and serve as a driving force for future research in that area. Moreover, the review facilitates the emergence of novel viewpoints that have not been

previously explored in literature and can have a significant impact on shaping practical applications and the future direction of the discipline (Torraco, 2016, p. 67). This study used integrative review for detailed comprehension of existing literature in the domain of algorithmic bias. The study also provides a novel point of view for existing empirical studies in algorithmic bias literature. Finally, a framework is developed which will provide an understanding of algorithmic bias in various AI capabilities.

2.1 Literature search -

This study focuses exclusively on the topic of algorithmic bias, and it includes the literature on gender bias, ethnic bias, any bias based on demographic differences, or consistent errors made by an algorithmic system. Since the scope of literature on algorithm bias is very broad, an exhaustive investigation was conducted on the topic "algorithmic bias", "algorithmic fairness", and "algorithmic discrimination." Additionally, for existing literature, sector-specific search pertaining to software and algorithms was conducted in various domains, such as those related to "policing", "judicial system", "social media", "chatbots", "voice assistants", or "education." Further, a search was conducted on the topics of "algorithm regulations" and "algorithm legislation" in order to explore the measures implemented for governing the use of algorithms. Moreover, this study included studies that were focused on mathematical strategies for addressing "discrimination" using purely technical approaches. Although initially intention was to restrict the scope of the search to peer-reviewed journal articles. However, the exception was made by including conference papers being published in reputable platforms such as arXiv, but only if they were directly related to the study. Various studies addressing the issue of algorithmic bias have been published in high-quality journals, with a focus on the topic of algorithmic bias. Therefore, a comprehensive database consisting of these studies was compiled. However, it was difficult to select only high-quality journal articles for inclusion because of - the frequent reprinting of reports, or the publication of derivative articles. When encountering cases where the news article was used as a reference by academic research articles, those articles were also obtained from the source paper. The study also included reports from various organizations if they directly tackle the issue of algorithmic bias. The goal was to achieve a comprehensive and inclusive approach while maintaining a clear focus.

To minimize subjectivity in sample selection, the study incorporated additional search terms in the databases (David & Han, 2004). To fulfill this objective, a search was conducted on Google Scholar using the keyword "algorithmic bias". The initial 50 search results were organized according to their relevance, and reviewed, and their variant terms, including "AI bias", "Algorithmic fairness", "bias in Machine Learning", and "fairness perception of ADM," were incorporated into the search query. Moreover, in accordance with the aim of investigating the study in algorithmic bias, even prior to the official emergence of the term, scholarly research conducted from 2001 to 2024 was incorporated into the study.

The search was conducted in January 2024 using keywords as a search from peer-reviewed journals from Google Scholar (Narayanan, Zane & Kemmerer, 2011; Micelotta, Lounsbury & Greenwood, 2017).

The articles underwent a screening process resulting in a final count of 186 studies. Subsequently, authors proceeded to carefully review the titles and abstracts of each of these articles, categorizing them by affixing a tick mark to those deemed promising, a question mark to those whose inclusion remained uncertain, and leaving unmarked articles that were deleted from consideration. As a result, a total of 22 articles were removed from the analysis, while 164 articles were deemed relevant and retained for further examination. Further, the authors employ forward and backward citation chaining and examine additional studies to include them in the review (Webster & Watson, 2002). Therefore 76 additional articles were included in the final sample. A total of 240 journal articles meeting the following criteria were identified: written in English, published in a peer-reviewed journal or press, and about the subject areas of social sciences, arts, psychology, business, and management. The Appendix provides a summary of the main conclusions derived from both empirical and non-empirical papers.

3. A Comprehensive Review of Literature on Algorithmic Bias

In the past few years, there has been an increasing amount of scholarly work dedicated to the topic of algorithmic bias. It can be asserted with confidence that the subject matter is experiencing a notable surge of interest within scholar communities. The literature consists of a wide range of works across several fields and occupations and has reached a logical stage where it can be reviewed and can be integrated. This integration offers a new direction for scholars to take the subject forward. This section aims to provide a comprehensive synthesis of the existing body of literature on algorithmic bias, elaborating on the numerous aspects that have been examined in previous research.

3.1 Harms due to Algorithmic bias-

The engagement process with algorithmic bias broadly involves three important stakeholders: algorithm creator, algorithm user, and organization for which the algorithm was developed (Schwartz et al., 2022), and each of them plays a crucial role in the phenomenon of algorithmic bias. Harm imposed by algorithmic bias can be classified into two areas: harm to users, and harm to organizations (Cramer et al., 2019). Also, Schwartz et al. (2022) mentioned that harm inflicted by algorithmic bias can harm individual, organizational, and societal levels. Some Key harms of algorithmic bias are elaborated in Appendix C.

3.2 The ethical considerations around algorithmic bias-

Today, with emerging technologies, ethical concern is one of the biggest challenges. (Richardson et al., 2021; Stahl, 2021). In the context of algorithms, without considering ethical and moral considerations due to algorithmic bias, the conversation on algorithms will be incomplete. One should be careful while asserting claims of biases on algorithms, as algorithmic bias is a complex phenomenon. It requires concrete specification of norms and standards before claiming an algorithmic bias to be negative or pernicious bias (Danks & London, 2017).

Mittelstadt et al., (2016) identified six ethical concerns of AI, based on how data is processed by the algorithm, which produces outcomes and motivates user's actions. IEEE Global initiative on ethics was launched by IEEE Standards Association (IEEE SA), in April 2016. The initiative aimed to provide standards, certification, and consensus-building for the ethical implementation of new technology. IEEE P703 Standard for "Algorithmic bias Consideration", provides a framework to improve algorithmic decision-making fairness, which is being developed and deployed by various organizations (Koene, Douthwaite & Seth, 2018).

Akter et al. (2021) examines algorithmic bias in AI-driven customer management. Technology adoption raises the consumers' knowledge about their personal data is being captured, which diminishes trust and causes discomfort (Bandara et al., 2020). According to Frow et al. (2011), service providers use advanced and invasive technologies without understanding their strategic aims or unethical intentions, mistreating and exploiting consumers. Such techniques limit customer decision-making by manipulating and distorting information. Thus, harmful practices must be addressed to reduce the social, economic, and ethical impacts of algorithmic systems. Algorithms can also promote ableism and can produce disparate results harming disabled people. According to Moura (2023), this technology will become an inseparable part of society, acting as a normalizing agent, defining societies' shared values, and ethics, and allocating limited resources. Hence such algorithmic bias will be a threat for disabled people.

Algorithmic processing is the major source of bias in current autonomous systems. This led to the rise of "ethical governors", which can modify algorithmic output, for autonomous systems are more likely will make ethical decisions, even if it reduces the likelihood of success. For instance, autonomous weapons may be provided with an ethical regulator, so they will not fire on perceived enemies if they are around UNESCO-protected historical sites (Danks & London, 2017).

3.3 Algorithmic Bias in AI Lifecycle-

According to Schwartz et al. (2022), to understand and mitigate bias in an AI system, it is necessary to understand the structure of how bias is present in an AI system. Organizations that are designing and developing AI keep track of the AI lifecycle for providing high-performing functionality but don't keep track of harm and manage it. Algorithmic bias in the AI lifecycle can be viewed as a four-step process, starting with the creation of a biased algorithm, spreading algorithmic bias, algorithmic bias detection, and algorithmic bias mitigation. Each of these steps is elaborated below.

3.3.1 Creation of biased algorithm-

It is vital to understand how and where in the operation of software, bias may get introduced so decision-making machinery will not reproduce the bias directly from the social world (Silva & Kenney, 2018). Silva & Kenney, (2018), mentioned nine kinds of biases enter an algorithm in different phases: (1) training data bias and algorithmic focus bias (input phase); (2) algorithmic processing bias (algorithmic operation phase); (3) transfer context bias, misinterpretation bias and automation bias (output phase); (4) non-transparency bias and consumer bias (user phase), and (5) feedback loop bias (feedback phase). These

papers provided structure and clarity about the concern and concept of algorithmic bias. Whereas, Schwartz, Vassilev, Greene, Perine, Burt, & Hall (2022) categorize AI bias into three dominant categories – Systemic bias, Human bias, and statistical & Computational bias. Third, Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, (2021), categorize the bias into groups of loops, such as (i) Data to Algorithm; (ii) Algorithm to User; and (iii) User to Data. Due to this loop phenomenon, definitions of bias get intertwined

3.3.2 Propagation of Algorithmic Bias-

The propagation of algorithmic bias has not gained significant attention in the literature, focus is limited to a few contexts. For instance, an algorithm organizes the user's flow of content to increase user engagement on social media platforms. So algorithmic bias could lead the user to interact with content that confirms to their belief (confirmation bias). Such algorithms create a polarizing feedback loop that reduces the diversity of the user's experience (filter bubble, echo chamber). People also experience social pressure in digital interaction, such as, the minority will adopt the majority opinion (Pansanella, Rossetti & Milli, 2022). Therefore, phenomena such as confirmation bias, filter bubbles, personalized recommendations of people, and peer pressure combined led to the creation of a system, where feeds received by users is biased, causing the dissemination of algorithmic bias.

Also, on social media, the demographics of influencers also impact the performance of influencers, for instance white, attractive, heterosexual people will enjoy the most privilege on social media, due to algorithmic visibility (Duguay, 2019; Stevens, 2021), over creators of color, LGBTQ creators, and plus size creators. Also, their content gets unfairly moderated or banned based on who they are (Ibrahim, 2010; Roth, 2015). Therefore, algorithmic visibility led to the dissemination of algorithmic bias (Maddox, 2022).

Even on an e-commerce platform such as Amazon, in the user's personalized account presence of filter bubbles was identified, where recommendations of misinformative health products contain more health misinformation as compared to neutral products. The personalized account was built from the user's search, click, marked top review, following contributors, and search on third-party websites (e.g., Google) (Juneja & Mitra, 2021). The study by Juneja and Mitra (2021), found that people who viewed misinformative products ended up purchasing them. Therefore, it pushed more misinformative items to a user who clicked on it, creating a problematic feedback loop. Therefore, on personalized accounts the formation of a filter bubble created by the platform and problematic feedback loop lead to the dissemination of algorithmic bias.

3.3.3 Detection and Mitigation of Algorithmic Bias –

Since fairness is a complex concept that depends on context and culture. Understanding the cause of bias is the first step in adopting effective algorithmic hygiene. Algorithms are increasingly becoming more and more complex. Researchers generally rely on algorithm outcome analysis to determine potential discrimination. Four strategies for identifying algorithmic bias are being identified by Fu, Huang, & Singh, (2020): Four-fifth (80%) rule, Regression analysis, Outcome test, and Benchmarking.

Also, Bellamy et al. (2019) in his study mentioned AI Fairness 360, which is an open-source toolkit to detect, understand and mitigate algorithmic biases. AIF360 aims to provide a deeper understanding of fairness metrics and mitigation techniques. AIF360 initial Python Package implements techniques from 8 published papers of the broader algorithm fairness community. It consists of 71 bias detection metrics, 9 bias mitigation algorithms, and a metric explanation facility that helps customers to understand the meaning of bias detection results.

4. Evaluation of empirical studies on Algorithmic bias

This review of empirical studies on algorithmic bias revealed significant advancement on various fronts, but a significant amount of work is still remaining to enhance our understanding of the phenomenon. As represented in Figure 1, empirical studies in algorithmic bias can be reasonably distinguished in three themes:

First, based on Internal algorithmic characteristics, which include properties of the algorithm, for instance, data used, and AI algorithm (Cowgill & Tucker, 2019). Therefore, it consists of literature discussing data biases, and bias in the algorithmic process, such as, how variables are weighed, and decisions made by institutions. *Second*, the evaluation of algorithmic bias from the perspective of how an algorithm is interpreted by users, for instance, perceived algorithmic fairness (Shin, 2020). *Third*, evaluation of Algorithmic bias literature from the behavioural science aspect (Behavioural Algorithmic bias)-, for instance, few people blindly trust algorithmic output, whereas sceptics interrogate algorithms to unveil some form of biases in algorithms. We argue that these issues are insufficiently explored and demand further empirical research. First, the key finding of past literature will be elaborated and then we focus on an area that deserves empirical validation.

4.1 Results from Previous Empirical Studies –

(i) Internal Algorithmic Characteristics aspect of algorithmic bias literature-

Empirical studies ***on internal algorithmic characteristics of algorithmic bias*** consist of : (i) defining algorithmic fairness, (ii) demonstrating algorithmic bias, and (iii) mitigating algorithmic bias. These studies described the technical and firm-level issues which lead to algorithmic bias.

(i) Defining algorithmic fairness - First, empirical studies defining algorithmic fairness will be discussed in this section. Fairness is being defined for three different levels - individuals, groups, and subgroups (Mehrabi et al. 2021). A review by Mehrabi et al. (2021) provided a widely used definition of fairness, these definitions were driven through empirical studies. Fairness definition at various levels are as follows: at the group level: Demographic parity (Dwork et al. 2012; Kusner et al., 2017), Conditional statistical parity (Corbett-Davies et al., 2017), Equalized odds (Hardt et al., 2016), Equal opportunity (Hardt et al., 2016), Treatment equality (Berk et al., 2021) Test fairness (Chouldechova, 2017); at subgroup level: Subgroup fairness (Kearns et al., 2018; 2019); and at individual level: Fairness through

unawareness (Grgic-Hlaca et al., 2016., Kusner et al., 2017) Fairness through awareness (Dwork et al., 2012) Counterfactual fairness (Kusner et al., 2017) (Mehrabi et al. 2021).

(ii) Demonstrating algorithmic bias - Second, empirical studies focused on algorithmic bias focused on demonstrating the presence of bias in an algorithmic system is being discussed. Amini et al. (2018), developed a Deep Neural Network (DNN) to address training data imbalance and potential bias in autonomous driving systems. Kay et al. (2015) demonstrated the presence of gender bias in image searches related to various occupations. Various studies have exhibited the presence of gender bias and racial discrimination in ads shown to users in online systems (Datta et al., 2014; Sweeney, 2013). Study by Angwin, Larson, Mattu, & Kirchner (2022) found that algorithms for predicting repeat offenders in the criminal justice system tend to discriminate on the basis of race. Torralba & Efros (2011) conducted a comparison study on popular image datasets and evaluated them on various criteria, such as relative data bias, cross-dataset generalization, and so on. Schmidt (2015), identified the presence of bias in word embedding. Caliskan, Bryson, & Narayanan, (2017) demonstrated that text corpora are the exact imprint of our historical bias, be it being morally neutral for flowers, and insects, problematic for gender/race, similar distribution of gender with careers and first name. Also, some studies demonstrated the presence of algorithmic bias in digital platforms, for instance, Lambrecht, & Tucker (2019), explored gender bias in ads provided by algorithms on Facebook, related to jobs in science, technology, engineering, and math fields; Dash et al. (2021), found that on Amazon platform sponsored recommendation were biased toward Amazon private label products; Xie, Yang & Yu (2021), elaborated the algorithmic bias in news recommendation in China's digital media; Park, Yu & Macy (2023) claimed that on Airbnb selection of same race endorsement by the consumer is due to top searches provided by recommender system and not due to content of the recommendation. Papakyriakopoulos & Mboya (2023), demonstrated racial bias in Google searches and claimed that discriminatory algorithmic outcomes resulted because of the training data set, and attitude of firm owners, and the algorithm designer. Lin et al. (2023), demonstrated algorithmic bias in leading search engine autocomplete, where discrimination was based on race, gender, and sexual orientation.

Several algorithmic bias assessment tools are also developed, which can assess fairness in a system. For instance, Saleiro et al. (2018) presented Aequitas, which is an open-source fairness and bias audit toolkit, that lets users test models for different biases and fairness matrices for individual/group/subgroups. It provides reports that help machine learning researchers, data scientists, and policymakers make informed and fair decisions for deployed algorithmic systems. Another toolkit is AI Fairness 360 (AIF360), which is developed with the aim of facilitating fairness in algorithms to be used in industrial settings and developing a framework for researchers to enable evaluation of algorithms. Its package consists of various fairness metrics and their explanations, and algorithms to mitigate biases in models and datasets. It also has an interactive web experience, to facilitate practitioners, and data scientists to implement the most appropriate tool while searching for solutions or in their work product (Bellamy et al., 2019).

(iii) Mitigating algorithmic bias - Third, the focus is made on empirical studies on algorithmic bias which discuss the mitigation of bias in an algorithmic system. Amini et al. (2019) targeted the biased data set having under-representation of a segment of society and developed an algorithm which mitigates potentially unknown and hidden biases in training data. The study by Bellamy et al. (2019), discusses about toolkit AIF360 can mitigate bias present in models and datasets. Various studies developed a method to mitigate bias borrowed from Saleiro et al. (2018) are : Hardt et al. (2016); Kamishima et al. (2011); Feldman et al. (2015); Kleinberg et al. (2016); Corbett-Davies et al. (2017); Zafar et al. (2017); Kearns et al. (2019); Noriega-Campero et al. (2019). A review study by Mehrabi et al. (2021), mentions several empirical studies in different domains (such as regression, clustering, natural language processing, and so on) to combat unfairness and bias in AI in order to attain fairness. Basically, methods to mitigate algorithmic bias lie in three categories- pre-processing, in-processing, and post-processing (Mehrabi et al., 2021; Schwartz, et al., 2022). Understanding fairness presence or bias in algorithmic systems will help people understand how and where an algorithmic bias can affect users and systems, and help researchers to identify potential points where biased or discriminating outcomes by algorithm will have negative consequences.

(iv) Attitude of firm owners, and the algorithm designer - Groves et al. (2024), discussed the failure of the algorithmic auditing system implemented in New York City, in July 2023, used for auditing automated employment decision-making tools (AEDTs). It occurs due to - narrow AEDT definition, flawed transparency-driven theory by law, industry lobbying, and challenges faced by auditors while data accessing. The study also provided recommendations for policymakers to develop a better algorithm auditing system.

(ii) Evaluation of algorithmic bias from consumer perception approach.

As AI is becoming widespread, it is important to address questions, such as how an algorithm is interpreted by users, and how users understand algorithm-based systems (Shin, 2020). Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2022), provided a systematic literature review of empirical studies on the perception of algorithmic fairness, the study included 58 empirical papers consisting of various disciplines and domains. Since, algorithmic decision-making, has some drawbacks: unfair algorithmic systems can systematically strengthen societal biases, causing marginalization of minorities, and without any restriction could harm certain society members (Veale and Binns, 2017; Žliobaitė, 2017). It is important because algorithms play an important role in various domains, such as interaction with business, education, government, and entertainment. Therefore, people are viewing the outside world from the algorithm's lens.

(i) FATE (Fairness, Accuracy Transparency, and explanation) – Transparency - Previous research studies were highly focused on the techno-centric approach, but now the focus has shifted to the user-centric design of algorithms. Public concern about bias in algorithmic systems discriminating based on race, gender, or other characteristics, also led to a call for transparency in problematic algorithmic systems (Goodman & Flaxman, 2017; O'neil, 2016; Weld & Bansal, 2019). For instance, a study by Springer &

Whittaker (2020), utilized an empirical method for understanding users' reactions toward transparent systems. Results tell that, initially user anticipated a transparent system to be better but changed their beliefs after experiencing a transparent system. Chen, Mislove & Wilson (2016), developed a methodology to detect algorithmic pricing, and empirically test its prevalence in the Amazon marketplace. The study explored the characteristics of algorithmic sellers and their strategies. The study aimed to increase transparency in such practices. Diakopoulos & Koliska (2017), conducted empirical research on algorithmic transparency in news media. Findings suggested the challenge of the human role in the adoption of algorithmic transparency, such as lack of incentive to organization and abundance of information to users. According to Lee et al. (2019), transparency and control over outcomes positively influence the fairness perception of the algorithm, as it includes humans in final decision-making. According to Datta et al. (2016), reporting transparency could have the potential for privacy breaches, and hence study explored the transparency-privacy trade-off. The result proved that transparency reports with the addition of little noise could be made private. Bujold, Parent-Rochelleau & Gaudet (2022), suggested that transparency of algorithmic surveillance positively affects procedural justice; and algorithmic performance management positively affects distributive justice; and procedural and distributive justice negatively influence the turnover rate.

FAT - Although algorithms have exhibited the potential to provide improved services to consumers, issues such as fairness, transparency, and accuracy (FAT) are intertwined with the algorithm's operation (Shin & Park, 2019). In this domain, several issues remain unsolved and problematic, for instance, whether the algorithm is fair or biased, the issue of assigning accountability in case of harmful outcome from the algorithm, and the issue of justifying goals, actions, and operations by the algorithm (Castelvecchi, 2018; Shin, 2019). Abdu et al. (2023), conducted content analysis of published articles on FAT, to identify in algorithmic fairness literature how race is being formalized and conceptualized. Results showed in algorithmic fairness literature racial categories are applied inconsistently and little explanation is provided for it. *And asked the algorithmic fairness community to re-examine the racial classification to align the field's intervention with its values.* According to Solyst et al. (2023), youths have a better capability to identify and articulate algorithmic bias. Young people, who have less awareness of technology and societal structure, can work with adults having better knowledge can lead to the development of fair and responsible AI. These studies will help in designing fair, transparent, and accountable AI, multidisciplinary teams will be required for implementing such a system (Turchi et al., 2024).

Explainable AI- Shin (2021) explored the explainable AI approach, in this study, 'causability' was conceptualized as an antecedent of explainability. The result suggested that including explanatory cues and cuasability, will increase users' trust in AI, as it brings transparency and accountability to AI. The study by John-Mathews (2021) talks about concerns about post-hoc explanations of black-box AI, a trend in explainable AI. Findings reported that post-hoc explanations have the tendency to provide biased information by algorithms' mechanisms or manipulate users to divert their attention.

(ii) *Algorithmic experience (AX)*: Due to the importance of algorithmic experience (AX), various empirical studies have been conducted to identify the process through which users create the perception of the algorithm. These studies provided guidelines for developers, to develop fair and responsible AI systems (Turchi et al., 2024).

(iii) *Trust*: The study by Shin et al. (2020), proposed an Algorithmic acceptance model as an analytic framework for human-algorithm interaction. As per result algorithmic experience is influenced by the user's understanding of fairness, transparency, and accuracy (FAT) and other experiences, which in turn are related to trust in the algorithm. According to the OECD (2019) and the European Commission (2019), trustworthy AI consists of four principles, among which fairness is one. However, it requires technical solutions to understand the societal implication of unfair algorithmic decisions (Barabas et al., 2020; Sloane and Moss, 2019), however various studies addressed when and why citizens perceive algorithmic decisions to be unfair.

Consumer fairness perception is necessary for implementing human-centric AI, as it informs developers to include ethical concerns for algorithmic systems to be implemented in the societal context (Kieslich et al., 2022). This could help to achieve a society-in-loop approach while designing an Algorithmic system (Rehwan, 2018). There is a need for more research work from theoretical as well as methodological stand from the context of non-Western countries, to develop a harmonized view on conceptualization and measurement of the algorithm's fairness perception.

(iii) Evaluation of Algorithmic bias literature from the behavioural science aspect (Behavioural Algorithmic bias)-

The third domain of empirical studies in algorithmic bias is which explores **algorithmic bias from a behavioural science aspect**, because of the human-in-loop approach in AI implementation (Schwartz et al., 2022). Therefore, AI decision-making can be affected by human biases, these biases are related to how a user perceives AI output while making the decision. AI is susceptible to these biases across the AI lifecycle once the AI application is deployed. There are various human biases, as they are a fundamental part of humans, the field around these biases is behavioural economics (Slovic & Tversky, 1982; Schwartz et al., 2022).

(i) *Automation Bias* – Automation bias occurs when users perceive the output of algorithms as being objective or factual. The user posits that the computer offers statistical computing that is devoid of bias or subjective influence. For instance, few people blindly trust algorithmic output (automation bias), whereas skeptics interrogate algorithms to unveil some form of biases in algorithms. There is inconsistency in the literature on automation bias. For instance, a study by Fritsch et al. (2022), using experimental method, showed that radiologists reading mammograms, supported by an AI-based system have a susceptibility to exhibit automation bias. They also mentioned the need to consider the effect of other such human-machine interactions. Whereas study by Alon-Barkat & Busuioc (2023), suggested that in the public sector bureaucrats are not prone to automation bias, but if the algorithmic output is as per group stereotypes (selective adherence bias), it may lead them to accept the decision.

The finding by Wright, Chen, Barnes & Hancock (2016), suggests that providing the reason for outcome improves performance and reduces automation bias; however, providing information that creates ambiguity increases complacency, which reduces performance and increases automation bias.

According to Horowitz & Kahn (2023) the knowledge, familiarity, and experience of AI, if lower likely causes algorithmic aversion and at an average level likely causes automation bias. Only after getting highly exposed to AI individuals become more balanced on whether to rely on AI decision aid or not. Jones-Jang, S. M., & Park, Y. J. (2023), explained two important psychological processes of how the failure of AI is evaluated by users. One mechanism when individuals have high expectations from AI's consistent performance causes 'automation bias'; and then they get frustrated by poor performance leading to 'algorithmic aversion'. Also, the study by Kim et al. (2023) mentioned that users' social identity can influence their perception of a biased algorithm, i.e., different social groups' lived experiences of discrimination influence their inference of biased algorithms. Schemmer, Köhl, Benz & Satzger (2022), stated that the impact of explanation on automation bias is dependent on the domain and kind of explanation. Explainable AI may increase dependency on AI, but if there is a confusing explanation, it either increases automation bias or leads to algorithmic aversion. Vered, Livni, Howe, Miller & Sonenberg (2023) conducted a study to design an explanation for an automated system. The study suggested that when automation bias is low no explanation is required to be offered, whereas when high performance is required and the willingness to accept automation bias, then there will be great use of explanation. According to Kupfer, Prassl, Fleiß, Malin, Thalmann & Kubicek (2023), individuals who received information of system error tend to have low automation bias, and individuals informed of their responsibility tend to have high automation bias. Also, there are other empirical behavioural studies in the algorithmic bias domain.

(ii) Interpretation Bias – Interpretation bias can be regarded as a mechanism via which users might introduce bias into an ostensibly neutral outcome. Bias occurs when individuals interpret an unclear output based on their personal internalized biases. For instance, a study by Lopez & Garza (2023) revealed that when users receive negative evaluation, they report high evaluation fairness for the human evaluator v/s AI evaluator, on the other hand, if users receive positive evaluation, they report a statistically insignificant difference in evaluation fairness level of user v/s AI. There is a lack of studies in behavioural science about algorithmic systems.

(iii) Feedback loop Bias - A notable characteristic of computation-based systems is their ability to generate additional data through all user activities. The algorithm is acquiring knowledge through the analysis of user behaviours. The study by Agan et al. (2023), focused on social media algorithms, suggests that the algorithm also fails because many a time, the choices made by users, deviate from their actual preferences, as the choices made by users are based on some specific context (known as automatic behaviour). These biases creep into the system, which can lead to biased behaviour by the algorithm which the user did not intend to, therefore automatic behaviour by the user leads to algorithmic bias. Therefore, there is a need to explore the behavioural aspect of AI from the user perspective since AI is going to be an integral part of the working environment and daily life of

individuals. This understanding will help organizations provide enhanced services and facilities to users and will help employees make better use of AI services.

The Table of literature of empirical studies is presented in Appendix A.

4.2 Need for more empirical studies in the future –

Now, this section examines the various facets and highlights the requirements of empirical studies:

Numerous definitions and approaches to fairness have been proposed and explored in the existing literature. However, it is important to note that the study in this field still is not complete. Algorithmic bias and fairness still have numerous research opportunities (Mehrabi et al., 2021). This section describes challenges in research of fairness and opportunities for under-studied research to conduct the study. Challenges: There exist a few unresolved challenges in fairness literature which is required to be addressed. These are: (1) Synthesizing a fairness definition. From the Machine Learning perspective, various definitions of fairness are being proposed in the literature. The definitions cover a broad spectrum of cases and hence exhibit some degree of disparity in fairness conceptualization. Therefore, it is very difficult to understand whether one fairness definition is applicable under different fairness conditions. Combining all fairness definitions into a single definition is still an open research problem, it will make algorithmic system evaluation more comparable and standardized. Having a standardized definition of fairness will help to deal with the issue of incompatibility with some of the current fairness definitions. (2) Need to move to equity from equality. Most of the fairness definitions present in the literature focus on equality, i.e., to ensure equal quantity allocation of resources among all groups or individuals. Little focus is being placed on equity, which means the allocation of resources to groups or individuals for them to succeed (Gooden, 2015). Therefore, operationalization of the equity definition and studying whether it enhances or contradicts the existing fairness definition can be an attractive research direction. (3) Identifying unfairness. Provided fairness definition, it should be possible to detect the presence of biases or unfairness in the dataset. Efforts have been made to address this issue of data bias, by identifying instances of Simpson's Paradox in arbitrary datasets (Alipourfard et al., 2018), still, more attention is required to be put on issues of unfairness because of the presence of multiple definitions and the absence of a process to detect them (Mehrabi et al., 2021).

Apart from the definitional inconsistency of fairness in literature, there is a requirement to conduct an empirical study on users' perceptions of the algorithmic fairness domain. For instance, developing a framework of algorithmic fairness by including four primary fairness dimensions: distributive, procedural, interactional, and informational fairness, enhances understanding of human-computer interaction (Starke et al., 2022). Also, from the perspective of AI-employee integration in organizations also opens the avenue of empirical studies, such as, (i) Cognitive issues - How do decision-makers trust output provided by AI systems? What controls are required when AI provides abnormal results and needs human intervention? (ii) Relational issues - How can employees build trust with an AI system/robot? And (iii) Structural Issues - How can we reskill workers to work successfully with AI systems? What type of technological and relational training is needed for nontechnical employees working with AI systems?

(Makarius et al., 2020). From the perspective of automation bias, there is a need for a deep understanding of the impact of explanation on automation bias. For instance, conducting mediation analysis and structural equation modeling to investigate the mediation effect. Such understanding will help to develop better human-AI collaboration (Schemmer et al., 2022).

Various literature reviews and conceptual studies have mentioned multiple research gap in the field of algorithmic bias. The study by Kordzadeh & Ghasemaghaei (2022) made recommendations for future information systems research they were – first, to examine the mechanism which influences user’s behaviour towards algorithmic bias when users experience algorithmic bias. Second, there is a need to understand how and within which circumstances algorithmic bias influences user behaviour towards outcomes provided by the algorithm, for instance, individual characteristics, task characteristics, technology characteristics, organizational characteristics, and environmental characteristics. These propositions majorly focus on the user’s behaviour and can help determine the fairness perception of the system developer. Whereas Puntoni, et al., (2021) acknowledged that AI technology is being embedded in various products which could impact consumers’ experience, hence the following question arises- Does difference in users’ awareness of biased algorithms make users feel misunderstood by AI? How does wrong social classification by AI output affect the choice and behaviour of consumers? How does the user decide which variable is used by AI for providing personalized recommendations? Which kind of classification leads consumers to feel misunderstood? Does the nature of the task impact the likeliness of feeling misunderstood? Also, Lopez & Garza (2023) mentioned that researchers can work to identify whether a negative judgment for “complex” tasks will influence users’ perception of fairness for human’s V/s AI.

Starke et al. (2022) mentioned the lack of studies from non-Western contexts, so researchers must focus on the impact of algorithmic bias among users from various countries. Also, there is a need for more theoretical and methodological groundwork to develop synchronized concepts and measurements of algorithmic fairness. Also, researchers should focus on interdisciplinary research to provide empirical evidence to - developers for the proper designing of algorithms; and to decision-makers for implementing algorithms following a society-in-loop framework. There is a requirement to integrate these three aspects of the study, i.e., *internal algorithmic characteristics*, which consists of technical solutions and goals defined by organizations for AI development; *user AI fairness perception*, which discusses the four aspects of AI design fairness, accountability, transparency and explainability (FATE); and human *behavioural science* which describe the human biases while interacting with smart systems. Combining these three aspects will provide the framework for designing an algorithm, as shown in Figure 2.

5. Future work on Algorithmic bias from behavioural aspect

There is a lack of study from the behavioural science perspective of algorithmic bias, literature majorly discussed automation bias and, to some extent interpretation bias by users. Having the understanding of human biases causing harmful algorithmic decision-making is still required, the framework provided in

this section will help researchers to provide an understanding of the phenomenon. It is important because AI is being developed from a Human-centric view, because humans and AI will be required to work together, i.e. human-in-loop approach (Schwartz et al., 2022).

A framework proposed in this section, as depicted in Fig. 3, is built on the basis of reviews and identifies gaps in existing published studies. The theoretical framework presented in this study is developed by integrating the concept of algorithmic bias sources present at different stages in the AI ecosystem (Silva & Kenny, 2018); and different AI capabilities (Puntoni et al., 2021). Different stages are involved in the AI ecosystem- data collection and storage, statistical and computational techniques, and output system- that empower products and services to carry out tasks that traditionally required human intelligence and now make decisions autonomously (Agrawal, Gans, and Goldfarb 2018). Framework by Puntoni et al. (2021), reflected that consumers engage with four AI capabilities, these are, data capturing capability, classification capability, delegation capability, and social interaction capability. The theoretical foundation and fundamental elements of the framework will be discussed in this section. Integrating the two ideas will provide a new direction to Algorithmic bias literature. This Framework on algorithmic bias is a novel contribution to literature.

The importance of the algorithmic bias domain is particularly because, today most of digital platforms and software algorithms are capturing more and more information to make informed decision-making (Gillespie, 2014; O'Neil, 2017) are ambiguous in society. Moreover, the advancement of digital technologies has led to increased complexity, as they have been increasingly integrated into social and economic decision-making processes in user's daily life as well. This integration occurs through either direct decision-making by algorithm or by generating outputs that influence human decision-making. This framework focuses on the phenomenon where output by AI influences human decision-making, because AI is getting implemented in organizations or available to users having humans in the loop, therefore it is required to develop an AI system with a human-centric approach where output by AI is utilized by the human user for decision making.

Scholars have reported that research on algorithmic bias on digital platforms is under-researched (Silva & Kenny, 2018). The underdevelopment can be attributed to the established practice of exploring algorithmic bias in high-impact domains, such as healthcare, policing, criminal justice, and so on. Understanding algorithmic bias in software and digital platforms is particularly important as they tend to structure social activities (Barley 2015; Scott and Orlikowski 2012).

This framework is developed by examining existing literature of human biases and human activities which lead to algorithmic biases that can harm online social and economic decision-making. For that purpose, the focus is to examine different biases and errors by humans while interacting with various AI capabilities. The study provides a classification and summary of the current body of literature on how AI capabilities could be subjected to reproduction, accentuation, or the creation of algorithmic bias. This study is elaborated upon the model proposed by Silva and Kenny (2018) and Puntoni et al. (2021), and can be generalizable for other biases and other AI capabilities. However, it is likely that each individual

set of biases and AI capabilities manifests themselves differently or concentrates in particular forms. The primary aim of this framework is to motivate researchers to build up conclusions outlined in this research and apply them in various specific circumstances in different digital platforms. Potential Sources of Bias in Algorithmic Processes as provided by Silva and Kenny (2018) is being represented in Table 1.

Table 1
Types of Algorithmic bias (Silva and Kenny, 2018)

Type of bias	Description
1) Training data bias -	Bias is introduced into the software as a result of its training data.
2) Algorithmic focus bias-	Algorithmic focus bias occurs within the dataset itself. As a society, humans established some categories, for instance, race and gender. Biases can occur when available information is included or excluded in the algorithm.
3)Algorithmic processing bias-	Termed as “processing” bias since bias is embedded in the algorithm itself. Some sources of such bias are – weighted- variables (Danks and London, 2017); algorithms do not account for differences in cases.
4) Transfer Context Bias-	Bias may manifest subsequent to the algorithm's provision of an output. The placement of output in an improper or unanticipated context has the potential to result in decisions that are discriminatory or biased.
5)Interpretation Bias-	Interpretation bias can be regarded as a mechanism via which users might introduce bias into an ostensibly neutral outcome. Bias occurs when individuals interpret an unclear output based on their personal internalized biases.
6) Non-Transparency of Outcomes Bias-	The dispersion of machine learning, along with the use of extensive databases containing numerous factors and the continuous vulnerability of algorithms, has led to a scenario wherein the explanations for outcomes are progressively becoming less transparent (Knight, 2017).
7) Automation Bias-	Automation bias occurs when users perceive the output of algorithms as being objective or factual. The user posits that the computer offers statistical computing that is devoid of bias or subjective influence.
8) Consumer Bias-	Consumer bias refers to the inherent bias that individuals may exhibit when engaging with digital platforms. The process efficiently transposes individuals' biases from the physical realm to the digital environment.
9) Feedback Loop Bias-	According to Zuboff (1988), a notable characteristic of computation-based systems is their ability to generate additional data through all user activities. The algorithm is acquiring knowledge through the analysis of user behaviours.

The coexistence of several biases in algorithmic processes or digital platforms is not always mutually exclusive, since it is possible that a system has multiple sources of biases that can interact in an intricate and complex manner, making them difficult to understand. Additionally, specific domains of activities (such as criminal justice and education), technology (facial recognition and search algorithms), or in some cases specific organizations (such as Uber, Airbnb, etc.) have also been explored. This study

provides information about the platforms and algorithms, the domains which are getting a lot of attention. The subsequent section examines the six sources of algorithmic biases occurs due to human biases or human activities; with four of AI capabilities being identified by Puntoni et al. (2021). The first three sources of biases, i.e., training data bias, algorithmic focus bias, and algorithmic processing bias occur because of technical or computational issues (Danks & London, 2017). Therefore, all AI capability will be affected by these three biases. Also, plenty of studies have focused on these three sources of biases (Schwartz et al., 2022). On the other hand, the remaining six sources of algorithmic bias, which by nature are based on human behaviours, are underexplored. Hence this framework focuses only on behavioural algorithmic bias sources -Transfer Context Bias, Interpretation Bias, Non-Transparency of Outcomes Bias, Automation Bias, Consumer Bias, and Feedback Loop Bias.

Types of AI Capabilities-

Second, this framework has also utilized the model provided by Puntoni et al., (2021). The study by Puntoni et al., (2021) bridged the two perspectives – one, value AI technology embeds in products and services for serving customers, and second, included sociological and psychological costs experienced by consumers while interacting with AI. The authors in this study identified four categories of AI capabilities: (1) data capture, (2) classification, (3) delegation, and (4) social interaction. Their study also accepted the conceptualization of AI as an ecosystem consisting of three fundamental components- data collection and storage, statistical and computational techniques, and output systems. These elements empower products/services to perform tasks traditionally considered to require intelligence and autonomous decision-making (Agrawal et al., 2018). These AI capabilities can also be understood as – listening (Capturing data), predicting (Classification), producing (Delegation), and communicating (Social Interaction). Consumer-AI experience as described in the framework provided by Puntoni et al., (2021) is shown in Table 2.

Table 2
Consumer -AI experience (Puntoni et al., 2021)

AI Capabilities	Description	Platforms Examples
Data Capture	The ability of AI system to listen, allows it to gather information regarding consumers and their surrounding environment.	All AI-enabled platforms
Classification	Organizations utilize the predictive capability of AI to provide highly personalized recommendations and services, aiming at maximizing customer engagement, satisfaction, and relevance. (Kumar et al., 2019).	OTT Platforms, Social Media, E-commerce
Delegation	A "delegation capability" refers to AI's capability to carry out tasks that would have otherwise been undertaken by users themselves. For instance, the Google Assistant.	Siri, Alexa, Google Assistant, Google map, ChatGPT
Social	The ability of AI to engage in reciprocal communication gives rise to what is commonly referred to as a "social experience".	Siri, Generative AI-Chatbots

Proposed Framework-

The framework related to algorithmic bias presented in this study differs from the models put out by Puntoni et al. (2021) in two significant ways. Previous research has examined the sociological and psychological dimensions of diverse AI capabilities. However, the present study aims to investigate the various biases that may occur when humans interact with different AI capabilities. Table 3 presents the summary of behavioural human biases with various AI capabilities. Secondly, this study discusses the consequences of these biases on users, businesses, and society (Appendix B).

Table 3
Summary of Literature review integrating Algorithmic bias with AI capabilities.

Consumer AI Capability	Type of Bias	Reference	Themes
1) AI-Data Capture Capability- (Puntoni et al., 2021)	Consumer biases	Greenwood et al., (2017)	Rating System, bias in user perception is feedback in the platform as data.
		Vincent, (2016)	Tay a Microsoft Chatbot on twitter – learned offensive responses from users’ interactions.
		Edelman and Luca, (2014)	Airbnb ♦ black hosts earn roughly 12% less “for a similar apartment with similar ratings and photos relative to [non-black] hosts”.
		Kakar et al. (2016)	Airbnb♦ Hispanic hosts’ - prices were 9.6% lower than those of equivalent non-Hispanic hosts, while Asian hosts’ listings had prices that were 9.3% lower.
		Stray, J. (2023).	Bias-producing feedback loops ♦ for a single user. The personalization process induces ♦ confirmation bias ♦ progressively poorer results.
	Feedback bias-	Greenwood et al., (2017)	Rating System, user evaluation is feedback in platform as data.
		Stray, J. (2023).	Recommenders are designed to respond to human feedback as a signal of relevance or quality. These loops can produce the positional, popularity, and polarization feedback loops.
		Silva and Kenney (2018)	Google search algorithm ♦respond to user’s query ♦ recorded and become input for succeeding searches, which improves the outcome in future search. ♦ algorithm is learning from user behaviour.
9) The AI Classification Capability – (Puntoni et al., 2021)	Non-transparency of outcome bias-	Datta et al., 2018.	User’s selection of ad setting preference has small effect on ad outcomes, rather individuals’ personal demographic characteristics and browsing history were major determinants of ads shown. “ad setting” do not provide full information regarding decision making.
		Rader and Gray (2015),	Judgement and belief of user about Facebook news feed algorithm presents.
		Mohseni, & Ragan, (2018).	and whether newsfeed algorithm is biased or not.
	Interpretation bias-	Mittelstadt & Russell, (2017)	Work in ML on explanations and interpreting model ♦ generate simple models. Basically,

Consumer AI Capability	Type of Bias	Reference	Themes
	Automation bias-		idea ◇ to create decision-making algorithm which accurately model the decision provided the current inputs ◇ various difficulties with explanations. In general, it is not clear whether models are interpretable by non-experts.
		Zarsky (2016)	Credit scores are fully automated, if process identifies a person have low credit score, they will lack access to credit, their score cannot improve, and hence they are trapped by algorithm
		Jackson et al. (2017)	Criminal justice system, algorithm provides risk assessment for making decision regarding bail or sentencing. Judges may give preference to computer generated recommendation compared to human based assessment
	Transfer context bias	Bode & Vraga (2018),	People tend to give more importance to information produced by algorithm or automation, known as automation bias, lead user to “over-accept” information they receive from computers- which also include information generated by Facebook related stories.
		Gallagher (2005)	Using credit score as variable to estimate job performance of employee
12) The AI Delegation Capability- (Puntoni et al., 2021)	Non-transparency of outcome bias-	Shang, Feng & Shah (2022),	one participant ◇ looking for explanation for Facebook advertisement ◇ it tells “you’re seeing this because the brand is trying to reach females ages 18 and up and people who live in the United States. [...]”. Which was super vague, and participant believe it as lie, and advertisers are not randomly targeting the person who lives in U.S.
		Natatsuka, Iijima, Watanabe, Akiyama, Sakai & Mori, (2022).	Application developers operates the VA (Voice- Assistance) application’s complex processing running on server. Therefore, behaviour of VA application is uncertain from user’s view. Although, Google Assistant directory ◇ provide some hint but did not provide accurate and detailed information about specific data the application access and process.
	Interpretation bias-	Rabassa, Sabri & Spaletta (2022)	lack between query and proposed offer by voice assistant, lead to fear to receiving biased offer by consumers. One consumer stated that she prefers to use Samsung voice assistant compared to Amazon’s Alexa, because it is perceived Alexa

Consumer AI Capability	Type of Bias	Reference	Themes
	Automation bias-		will primarily recommend Amazon's products and brand
		Rabassa, Sabri & Spaletta, (2022)	some consumers were blindly trusting the algorithm choice made by voice assistant. They believe that voice assistant is faithful towards customer and increasing their welfare by choosing product that best suit their needs.
		Bode & Vraga (2018),	There is a tendency among individuals to assign greater significance to information generated by algorithms or automation, a phenomenon commonly referred to as automation bias. These bias leads user to excessively accept and rely upon information received from computer systems, including the content provided by Facebook through news feed.
13) The AI Social Capability – (Puntoni et al., 2021)	Non-transparency of outcome bias-	Khurana et al. ((2021)	Chatbot appears as “black-box” to user, which make it difficult to understand why something is not working, what could be done to recover from breakdown. Lack of transparency, impacts user's perception of usefulness and trust in system
	Interpretation bias-	Brahnam and De Angeli, (2012)	Female representing chatbot was subjected to more implicit and explicit sexual attention and swear word compared to male-presenting chatbot. if avatar was presented as black adult, it often faced racist attacks
		Marino (2014)	chatbot developed for answering questions about Caribbean Aboriginal culture, chatbot was represented as a Caribbean Amerindian person, which lead to unintended stereotype because user perceived chatbot's behaviour as standard for people from the represented population.
	Transfer context bias-	Adiwardana et al. (2020)	Some weaknesses in open-domain chatbot, such as, they often reply in manner which is vague and generic. Response by chatbot could be considered as sensible, if it is specific to given context. For instance, if A says, “I Love Tennis”, and B responds “that's nice”,
		Shuster, Poff, Chen, Kiela, & Weston, (2021)	Knowledge is implicitly stored in weights of large language models, consisting of billion parameters which make it possible for the agents to speak knowledgeably on open-domain topics. But, unfortunately even largest model suffers from “hallucination” problem

Consumer AI Capability	Type of Bias	Reference	Themes
			(Maynez et al., 2020) where the response is factually incorrect statement. They often mix facts between two similar entities, and even make error when one token being incorrect is difference between right and wrong.

Therefore, AI data-capturing capability is prone to consumer bias and feedback loop bias. And AI’s Classification capability, Delegation capability, and Social Capabilities are prone to non-transparency of outcome bias, interpretation bias, transfer of context bias and automation bias an depicted in Table 3.

A summary of the Literature review integrating Algorithmic bias literature review with AI capability is presented in Appendix B, and the list of research questions obtained from the framework is mentioned in Table 4.

Table 4
List of Research Questions obtained from Framework-

AI-Data Capturing Capability-
Q1 Whether the particular AI system used by the users is prone to Consumer bias?
Q2 Whether the particular AI system used by the users is prone to Feedback Loop bias?
Q3 Is consumer bias in AI systems is affecting consumer experience?
Q4 Is consumer bias in AI systems is having a harmful impact on society?
Q5 Is feedback loop bias in AI systems is affecting consumer experience?
Q6 Is feedback loop bias in AI systems is having a harmful impact on society?
Q7 How does Consumer bias and Feedback loop bias can harm organizational decision-making?
AI-Classification Capability
Q1 Whether the particular AI having classification capability used by the users is prone to non-transparency of outcome/Interpretation/Automation/Transfer of Context bias?
Q2 Whether the AI with Classification Capability possessing non-transparency of outcome/Interpretation/Automation/Transfer of Context bias have a negative or positive effect on consumer experience?
Q3 Whether the AI with Classification Capability possessing non-transparency of outcome/Interpretation/Automation/Transfer of Context bias have a harmful impact on society?
Q4 How does non-transparency of outcome/Interpretation/Automation/Transfer of Context bias in AI having classification capability can harm organizational decision-making?
AI-Delegation Capability
Q1 Whether the particular AI having delegation capability used by the users is prone to non-transparency of outcome/Interpretation/Automation/Transfer of Context bias?
Q2 Whether the AI having Delegation Capability possessing non-transparency of outcome/Interpretation/Automation/Transfer of Context bias have a negative or positive effect on consumer experience?
Q3 Whether the AI having Delegation Capability possessing non-transparency of outcome/Interpretation/Automation/Transfer of Context bias have a harmful impact on society?
Q4 How does non-transparency of outcome/Interpretation/Automation/Transfer of Context bias in AI having delegation capability can harm organizational decision-making?
AI -Social Capability
Q1 Whether the particular AI having social capability used by the users is prone to non-transparency of outcome/Interpretation/Automation/Transfer of Context bias?
Q2 Whether the AI having Social Capability possessing non-transparency of outcome/Interpretation/Automation/Transfer of Context bias have a negative or positive effect on consumer experience?

AI-Data Capturing Capability-
Q3 Whether the AI having Social Capability possessing non-transparency of outcome/Interpretation/Automation/Transfer of Context bias have a harmful impact on society?
Q4 How does non-transparency of outcome/Interpretation/Automation/Transfer of Context bias in AI having Social capability can harm organizational decision-making?

6. Potential Research gaps

This review of algorithmic bias literature provides a holistic view of the interdisciplinary nature of algorithmic bias research. However, various research gaps in the literature are required to be addressed for the advancement of research on this topic. This section examines various gaps and limitations and provides recommendations for future research.

6.1 Definitional Concern-

Literature on algorithmic bias has a definitional issue. The term “bias” has overlapping meanings which makes discussion on the topic of algorithmic bias highly complicated (Crawford, 2017; Blodgett et al., 2020). For instance, Blodgett et al., (2020) conducted a survey on bias in NLP and identified several issues to be clarified, such as the manner author defined the algorithmic bias, and how exactly the system is biased. They mentioned that various studies do not have the proper motivation, several times it remains unclear in what manner “bias” is harmful and for whom, and many a time paper targets on the narrow source of algorithmic bias. This led to a rise in potential research questions in three domains.

Believes and social hierarchies-

While analyzing algorithmic bias it is important to engage with literature on “bias” in the social domain, which considers the relation between users' beliefs and social hierarchies. Disciplines such as sociology, and social psychology study social interactions, their impact on individuals, how individuals behave in society, and so on. This led to the following research questions:

- How do social hierarchies and beliefs influence the decision while developing and deploying algorithmic systems? What type of algorithmic system is developed due to these decisions and what kind of system are ruled out?
- How do Algorithmic systems reproduce or transform social/individual beliefs? Which kinds of beliefs and practices are considered as good or bad? Could good beliefs mean ideologies that could be easily handled by an algorithmic system?
- Which of these representational harms are being measured or mitigated? Whether these are the greatest moral concerns or just these concerns can be effectively addressed by algorithmic fairness techniques. Are there other harms that is required to be analyzed?

Conceptualizations of “bias”-

Studies analyzing “bias” in the algorithmic system must provide arguments explicitly for why the system behaviour considered as “bias” is detrimental, how and for whom, they must also provide normative reasoning for provided arguments. That is, researchers must formulate a proper conceptualization of “bias”. Therefore, arises following questions:

- Which system behaviour can be considered as “biased”? What could be the potential sources, (for instance, data, general assumptions, and so on)? How such behaviour by the system is harmful, and why? Which of the social values supports the conceptualization of “bias”?

Users believe in practice-

This perspective stands on the recognition that the relation between users’ beliefs and the social system, provides multiple directions to examine users’ beliefs in practices. The researcher could focus on important elements. First, since different social groups have different lived experiences, researchers working on algorithmic bias must consider the lived experiences of community members. Second, it is required to investigate the power relation between these communities and technologists. Researchers proposed technical mitigation techniques to remove algorithmic bias – e.g., use new training data points for a better model – sustain this power dynamics by (a) assuming that algorithmic system will continue to exist, rather than questioning whether such system should be built at all, (b) technologists decides the development and deployment of such system (Bennett & Keyes, 2019; Cifor et al., 2019; Green, 2019; Katell et al., 2020). Hence following questions arise:

- How do social groups become aware of the Algorithmic system? Whether they resist against them, if yes, why?
- What additional losses does the community face due to the discriminatory behaviour of the algorithmic system?
- Does an algorithmic system shift power in the hands of oppressive institutions (for instance, predicting that a particular group does not want, unfair allocation of resources and opportunities, surveillance, or censorship), or take it away from such organizations?
- Who is responsible for the development and deployment of algorithmic systems? How do these decision-making processes maintain power relations between groups/individuals impacted by algorithmic systems and technologists? Whether such practices can be changed or these relations could be reimagined?

6.2 Requirement of Theoretical Underpinning -

A major concern in algorithmic bias literature is the lack of a comprehensive theoretical framework. Due to lack of a theoretical framework for guiding empirical research on algorithmic bias causes researchers to use various approaches for theoretically supporting their hypothesis. As illustrated in Table 5, theories in the fields of information systems, organization behaviour, and social psychology is being utilized by researchers studying algorithmic bias.

Table 5
Theoretical foundation of Algorithmic bias literature

Theoretical lens	How is the theory used in Algorithmic bias literature?	Examples from Algorithmic bias literature
Stimulus-organism-response theory	According to this theory, the internal (psychological) state of an individual gets influenced by environmental stimuli, which influences behavioural response (Mehrabian & Russell, 1974).	Kordzadeh & Ghasemaghaei (2022)
Organisational justice theory	As per this theory, from the employee's perspective justice means the "extent to which organization or its top management is perceived to act consistently, truthfully, respectfully and equitably while making a decision" (Colquitt & Rodell, 2015).	Kordzadeh & Ghasemaghaei (2022)
Confirmation bias theory	This theory posits that human tends to perceive technologies and information positively which confirms to their preexisting beliefs (Jussupow et al., 2020; Kahneman, 2011).	Kordzadeh & Ghasemaghaei (2022)
Anchoring bias theory	According to this theory, while making decisions people tend to weigh heavily the first piece of information they have.	Kordzadeh & Ghasemaghaei (2022)
Dual process and dual-system theories	This theory posits that people tend to use two sets of decision-making processes, for instance, the need for cognition – consciously or unconsciously. IS researchers can study whether the user's characteristics influence their detection and reaction to algorithmic bias.	Kordzadeh & Ghasemaghaei (2022)
Socio-technical theories	According to socio-technical theory, every organization consists of two subsystems, technical and social subsystems. Fit between these subsystems is required to overcome the difficulties of workers, and for achieving expected benefits.	Schwartz, et al. (2022).
Consumer experience	Schmitt (1999) provided a multidimensional view and found out five kinds of experiences: sensory (sense), affective (feel), cognitive (think), physical (act), and social-identity (relate) experiences.	Puntoni et al. (2021)

Since AI systems are being established mostly in high-risk settings to deal with known biases and subjectivity of humans. Various questions still remain to be answered regarding the optimal configuration between automation and humans. Therefore, future research can build a bridge between technical communities with various subfields such as psychology, organization behavior, and human factors is necessary. This will help in developing formal guidance regarding implementing human-in-loop processes to reduce human, systematic, and computational bias (Schwartz, et al., 2022).

7. Discussion

This study demonstrated that algorithmic bias is a potential research stream, because of ubiquity of algorithm in society and their use in various domains in future. This integrative review makes a contribution by providing understanding of theoretical development opportunity in algorithmic bias domain. This study also provides guidelines for organizations and users, to be conscious of algorithmic biases while using AI tools. This section discusses the implication of this study for theory, policymakers and researchers in IS domain.

This study made three contributions towards in algorithmic bias. First, this study provides systematically organized and current state of existing research in algorithmic bias literature which help in defining the current outline of this literature, as it provided the three domain in which studies is being conducted in Algorithmic bias literature, i.e., first, Internal algorithmic characteristic, which include technology as well as thought process while developing an algorithm; second, user perceived fairness, which discusses how to improve fairness perception of an AI tool to enhance AI adoption among users; and third is behavioural aspect of AI, i.e., what are the behavioural implication of algorithmic system, for instance, automation bias, and still various other behavioural aspect is required to be found out in the algorithmic system. This study is different from previous reviews as it provides a holistic discussion of existing research in algorithmic bias topics.

Second, this study provides limitations and research gaps present in existing literature, and theme-specific research questions provide a solid foundation for researchers interested in studying the phenomenon of algorithmic bias. Third, this study proposed a framework that provides structure to under-researched algorithmic biases in current literature and provides an opportunity for future research. The Framework will also act as a guideline for organizations and users using algorithmic systems, to be cautious of various biases present in these AI systems and reduce the consequences due to it. Framework also provides guidelines for developers to keep things into consideration, such as fairness, transparency, and explainability of the AI system, as this will help in AI adoption, having various capabilities, such as classification, delegation or social. Also, the research gap provided by this framework will help in gaining novel insight of this phenomenon and will help to expand knowledge in this domain.

This study has major implications for policymakers. Firstly, it is imperative for policymakers to acknowledge the widespread utilization of AI technologies. From present regulations around AI and various challenges mentioned in this review regarding algorithmic bias policymakers can derive valuable insights to formulate regulations pertaining to the creation and dissemination of algorithmic bias, although the review suggests that it is difficult to identify perfect solution to control the potential threat which could occur due to algorithmic bias, but consistent calibration of policies is essential to reduce negative impact. Second, since AI have various positive impact study will provide direction to policymakers, as policymakers will make sure that implementing the regulation of AI do not leads to harm to the positive application of AI. Therefore, policymakers must ensure that the proposed regulation provides an advantage for the advancement of advantageous application of AI technology. Third, the study highlights the ongoing growth in the adoption of AI technology, emphasizing the need for

algorithmic bias detection mechanisms to remain aligned with advancements in AI technologies. This encourages policymakers to make sure to provide adequate incentives and encouragement, by making investments in developing solutions to detect and mitigate algorithmic bias. This review also provides insights to platform players to develop guidelines and standards for the governance of platforms.

Also, various research directions are offered in this review for researchers of the Information System (IS) domain. Three critical aspects have been described below. First, today diverse algorithms are prevalent in society, for instance, News-ranking algorithms, social media bots control the information visible to users; algorithms making loan decisions based on credit scoring; and online pricing algorithms provide different cost to different consumers (Rahwan et al., 2019), hence algorithms are influencing humans' decisions and experiences (Puntoni et al., 2021). This environment can cause harm to humans, since it can influence individuals to get engaged in toxic situations such as fake news, access to limited information due to the formation of filter bubbles, spending more time on social media, receiving unfair prices on e-commerce, and so on. Such incidents could lead to harmful experiences for users. These persuasive technologies creating new socio-technical environments open up multiple options for IS researchers to examine how these technologies are influencing user's emotions, behaviour, and cognitive abilities. Second, algorithmic bias raises various ethical concerns such as disparate treatment of groups/individuals, privacy concerns, and harmful outcomes by algorithms, despite that algorithmic systems have various benefits, for instance, algorithmic justice, policing, healthcare, and so on. Therefore, IS researchers have the opportunity to make substantial contributions by providing guiding principles for algorithm designers to reduce algorithmic bias in each stage and embrace the privacy of users. Third, there is a lack of clarity in the literature regarding understanding the impact of algorithmic bias and its enabling characteristics of spreading online. Algorithmic bias has the risk of exacerbating social prejudices and harming individuals, organizations, and society. Platform providers and consumers have to deal with negative consequences due to biased information and decisions from algorithms. To mitigate these harmful consequences a socially aware algorithm is essential which will be able to counter biased decisions (Schwartz et al., 2022). Therefore, under such circumstances IS researchers are in a favorable position to investigate social and technical elements. Their finding will be able to provide valuable insight to platform players and regulatory bodies, informing them of potential risks posed by algorithmic bias and their negative impact on consumers.

8. Conclusion and Limitations

This study consists of an integrative review of the literature on algorithmic bias. Integrative reviews serve the purpose of examining a particular topic in-depth and amalgamating existing studies to develop a novel insight and conceptual framework (Torraco, 2005; Webster & Watson, 2002). This study is one of the attempts which offer a comprehensive and critical review on the topic of algorithmic bias, and different from past reviews, which had having narrow focus. This study has conducted an assessment of the current body of literature on algorithmic bias and identified specific areas that require more investigation. Additionally, the framework we have proposed will serve as a valuable tool for organizing and conducting future research on algorithmic bias in a systematic manner. This review aims to

encourage interdisciplinary collaboration among academics and foster a comprehensive knowledge of algorithmic bias as a multi-domain phenomenon. By addressing the challenges posed by this significant technical innovation, researchers may effectively mitigate its negative consequences and harness its potential for providing value.

Limitations

Notwithstanding the offered contributions in this study, it is not devoid of limitations. First, this study is primarily grounded on secondary data obtained from academic research on algorithmic bias. Although the study attempted to include grey literature in this review, there exist a few limited sources, therefore future studies can make enhanced efforts and provide a comprehensive understanding of the topic. Similarly, we identified some papers that were inaccessible due to a paywall, therefore preventing us from accessing the complete text. Nevertheless, we endeavoured to capture the fundamental substance of these publications in our assessment. Second, for assistance in this integrative review, a set of contextual and relevant keywords along with commonly used databases were used, for the selection of relevant publications. Therefore, this study is completely original, thorough, and critical in nature, and was based on keywords, but the evolving character of these phenomena and the resolution of definitional issues may provide new opportunities in the future and include more studies in future literature review. Hence, this study will act as a platform for further research into the algorithmic bias phenomenon.

Declarations

Disclosure of potential conflicts of interest-

No potential conflict of interest was reported by the author(s).

Author Contribution

This study is conducted by a single author (Amit Kumar Chaudhary)-1) First Literature search was done on the topic of Algorithmic bias.2) Themes were conceptualized, and emerging themes were identified.3) Written the paper

References

1. ABC News (2020). Government concedes flaws but refuses to apologise for its unlawful. <https://www.abc.net.au/news/2020-05-31/robodebt-federal-government-christian-porter-no-apology/12304672>
2. Abdu, Amina A., Irene V. Pasquetto, and Abigail Z. Jacobs. "An Empirical Analysis of Racial Categories in the Algorithmic Fairness Literature." In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1324-1333. 2023.

3. Adiwardana, D., Luong, M. T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., ... & Le, Q. V. (2020). Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.
4. Agan, A. Y., Davenport, D., Ludwig, J., & Mullainathan, S. (2023). *Automating automaticity: How the context of human choice affects the extent of algorithmic bias* (No. w30981). National Bureau of Economic Research.
5. Agarwal, P. (2019, March). Redefining banking and financial industry through the application of computational intelligence. In 2019 Advances in Science and Engineering Technology International Conferences (ASET) (pp. 1-5). IEEE.
6. Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press.
7. Ahluwalia, Rohini, Robert E. Burnkrant, and H. Rao Unnava (2000), "Consumer Response to Negative Publicity: The Moderating Role of Commitment," *Journal of Marketing Research*, 37 (2), 203–14.
8. Aitken, M., Toreini, E., Carmichael, P., Coopamootoo, K., Elliott, K., & van Moorsel, A. (2020). Establishing a social licence for Financial Technology: Reflections on the role of the private sector in pursuing ethical data practices. *Big Data & Society*, 7(1), 2053951720908892.
9. Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y. K., D'Ambra, J., & Shen, K. N. (2021). Algorithmic bias in data-driven innovation in the age of AI. *International Journal of Information Management*, 60, 102387.
10. Akter, S., Michael, K., Uddin, M. R., McCarthy, G., & Rahman, M. (2022). Transforming business using digital innovations: The application of AI, blockchain, cloud and data analytics. *Annals of Operations Research*, 1-33.
11. Akter, Taslima. "Privacy considerations of the visually impaired with camera based assistive tools." In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, pp. 69-74. 2020.
12. Alipourfard, N., Fennell, P. G., & Lerman, K. (2018). Can you Trust the Trend?. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM.
13. Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2), 211-236.
14. Alon-Barkat, S., & Busuioc, M. (2023). Human–AI interactions in public sector decision making: "automation bias" and "selective adherence" to algorithmic advice. *Journal of Public Administration Research and Theory*, 33(1), 153-169.
15. Amini, A., Schwarting, W., Rosman, G., Araki, B., Karaman, S., & Rus, D. (2018, October). Variational autoencoder for end-to-end control of autonomous driving with novelty detection and training debiasing. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 568-575). IEEE.
16. Amini, A., Soleimany, A. P., Schwarting, W., Bhatia, S. N., & Rus, D. (2019, January). Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 289-295).

17. Andr e, Quentin, Ziv Carmon, Klaus Wertenbroch, Alia Crum, Douglas Frank, William Goldstein, et al. (2018), "Consumer Choice and Autonomy in the Age of Artificial Intelligence and Big Data," *Customer Needs and Solutions*, 5 (1/2), 28–37.
18. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2022). Machine bias. In *Ethics of data and analytics* (pp. 254-264). Auerbach Publications.
19. Arkin, R. C., Ulam, P., & Wagner, A. R. (2011). Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the IEEE*, 100(3), 571-589.
20. Arora, A. (2020). Conceptualising artificial intelligence as a digital healthcare innovation: an introductory review. *Medical Devices: Evidence and Research*, 223-230.
21. Artificial Intelligence Algorithm: Everything You Need To Know About It. (2021, June). Rock Content. Retrieved from <https://rockcontent.com/blog/artificial-intelligence-algorithm/>
22. Artificial solutions (2019). Why Chatbots Fail: Limitations of Chatbots. *voice-tech-podcast*. <https://medium.com/voice-tech-podcast/why-chatbots-fail-limitations-of-chatbots-7f291c4df83f>
23. Ashktorab, Z., Jain, M., Liao, Q. V., & Weisz, J. D. (2019, May). Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1-12).
24. Bajracharya, A., Khakurel, U., Harvey, B., & Rawat, D. B. (2022, October). Recent Advances in Algorithmic Biases and Fairness in Financial Services: A Survey. In *Proceedings of the Future Technologies Conference (FTC) 2022, Volume 1* (pp. 809-822). Cham: Springer International Publishing.
25. Baker, R. S., & Hawn, A. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 1-41.
26. Bandara, R., Fernando, M., & Akter, S. (2021). Managing consumer privacy concerns and defensive behaviours in the digital marketplace. *European Journal of Marketing*, 55(1), 219-246.
27. Barabas, C., Doyle, C., Rubinovitz, J. B., & Dinakar, K. (2020, January). Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 167-176).
28. Barati, M., & Ansari, B. (2022). Effects of algorithmic control on power asymmetry and inequality within organizations. *Journal of Management Control*, 33(4), 525-544.
29. Barley, S. R. (2015). Why the internet makes buying a car less loathsome: How technologies change role relations. *Academy of Management Discoveries*, 1(1), 5-35.
30. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California law review*, 671-732.
31. Barocas, S., Biega, A. J., Fish, B., Niklas, J., & Stark, L. (2020, January). When not to design, build, or deploy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 695-695).

32. Barrett, Lindsey. 2017. "Reasonably Suspicious Algorithms: Predictive Policing at the United States Border." *NYU Rev. L. and Soc. Change* 41, no. 3: 327-365.
33. Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics*, 143(1), 30-56. Available: <https://www.nber.org/papers/w25943>
34. Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5), 4-1.
35. Belluz., J., 2016. Amazon is a giant purveyor of medical quackery. <https://www.vox.com/2016/9/6/12815250/amazon-health-products-bogus>
36. Bembeneck, E., Nissan, R., & Obermeyer, Z. (2021). To stop algorithmic bias, we first have to define it. *Policy Commons*
37. Bennett, C. L., & Keyes, O. (2020). What is the point of fairness? Disability, AI and the complexity of justice. *ACM SIGACCESS Accessibility and Computing*, (125), 1-1.
38. Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1), 3-44.
39. Blier, N. (2019). Bias in AI and machine learning: Sources and solutions. *Lexalytics. August 15, 2019.* <https://www.lexalytics.com/lexablog/bias-in-ai-machine-learning>.
40. Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in nlp. *arXiv preprint arXiv:2005.14050*.
41. Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.
42. Brahnam, S., & De Angeli, A. (2012). Gender affordances of conversational agents. *Interacting with Computers*, 24(3), 139-153.
43. Brakus, J. J., Schmitt, B. H., & Zarantonello, L. (2009). Brand experience: what is it? How is it measured? Does it affect loyalty?. *Journal of marketing*, 73(3), 52-68.
44. Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. In *Internet Science: 4th International Conference, INSCI 2017, Thessaloniki, Greece, November 22-24, 2017, Proceedings 4* (pp. 377-392). Springer International Publishing.
45. Brandtzaeg, P. B., & Følstad, A. (2018). Chatbots: changing user needs and motivations. *interactions*, 25(5), 38-43.
46. Bridgeman, B. (2009). Experiences from large-scale computer-based testing in the USA. The transition to computer-based assessment, 39.
47. Broussard, M. (2018). Artificial unintelligence: How computers misunderstand the world. mit Press.
48. Bujold, A., Parent-Rochelleau, X., & Gaudet, M. C. (2022). Opacity behind the wheel: The relationship between transparency of algorithmic management, justice perception, and intention to quit among

- truck drivers. *Computers in Human Behavior Reports*, 8, 100245.
49. Bunt, A., Lount, M., & Lauzon, C. (2012, February). Are explanations always important? A study of deployed, low-cost intelligent interactive systems. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces* (pp. 169-178).
 50. Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society*, 3(1), 2053951715622512.
 51. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183-186.
 52. Castelvechi, D. (2016). Can we open the black box of AI?. *Nature News*, 538(7623), 20.
 53. Chen, J., Geyer, W., Dugan, C., Muller, M., & Guy, I. (2009, April). Make new friends, but keep the old: recommending people on social networking sites. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 201-210).
 54. Chen, L., Mislove, A., & Wilson, C. (2016, April). An empirical analysis of algorithmic pricing on amazon marketplace. In *Proceedings of the 25th international conference on World Wide Web* (pp. 1339-1349).
 55. Chen, R. J., Wang, J. J., Williamson, D. F., Chen, T. Y., Lipkova, J., Lu, M. Y., ... & Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, 7(6), 719-742.
 56. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2), 153-163.
 57. Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018, January). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency* (pp. 134-148). PMLR.
 58. Chowdhury, S., Dey, P., Joel-Edgar, S., Bhattacharya, S., Rodriguez-Espindola, O., Abadie, A., & Truong, L. (2023). Unlocking the value of artificial intelligence in human resource management through AI capability framework. *Human Resource Management Review*, 33(1), 100899.
 59. Christy, A., Gandhi, G. M., & Vaithyasubramanian, S. (2019). Clustering of text documents with keyword weighting function. *International Journal of Intelligent Enterprise*, 6(1), 19-31.)
 60. Ciampaglia, G. L., & Menczer, F. (2018). Misinformation and biases infect social media, both intentionally and accidentally. *The Conversation*, 20.
 61. Cifor, M., Garcia, P., Cowan, T. L., Rault, J., Sutherland, T., Chan, A., ... & Nakamura, L. (2019). Feminist data manifest-no. Cit. on, 119.
 62. Conroy, N. K., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology*, 52(1), 1-4.
 63. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017, August). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 797-806).

64. Cowgill, B., & Tucker, C. E. (2019). Economics, fairness and algorithmic bias. *preparation for: Journal of Economic Perspectives*.
65. Crain, M. (2018). The limits of transparency: Data brokers and commodification. *new media & society*, 20(1), 88-104.
66. Cramer, H., Garcia-Gathright, J., Reddy, S., Springer, A., & Takeo Bouyer, R. (2019, May). Translation, tracks & data: an algorithmic bias effort in practice. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-8).
67. Crawford, K. [The Artificial Intelligence Channel]. (2017). The Trouble with Bias - NIPS 2017 Keynote - Kate Crawford [Video]. YouTube. https://youtu.be/fMym_BKWQzk
68. Cruz, T. M. (2020). Perils of data-driven equity: safety-net care and big data's elusive grasp on health inequality. *Big Data & Society*, 7(1), 2053951720928097.
69. Cummings, W. (2018). Diamond and Silk tell Congress,'Facebook censored our free speech!'. USA Today. Available online: <https://bit.ly/3r6FsJp>.
70. Danks, D., & London, A. J. (2017, August). Algorithmic Bias in Autonomous Systems. In *Ijcai* (Vol. 17, No. 2017, pp. 4691-4697).
71. Dash, A., Chakraborty, A., Ghosh, S., Mukherjee, A., & Gummadi, K. P. (2021, March). When the umpire is also a player: Bias in private label product recommendations on e-commerce marketplaces. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 873-884).
72. Datta, A., Tschantz, M. C., & Datta, A. (2014). Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491*.
73. Datta, A., Sen, S., & Zick, Y. (2016, May). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)* (pp. 598-617). IEEE.
74. Datta, A., Datta, A., Makagon, J., Mulligan, D. K., & Tschantz, M. C. (2018, January). Discrimination in online advertising: A multidisciplinary inquiry. In *Conference on Fairness, Accountability and Transparency* (pp. 20-34). PMLR.
75. Davenport, T., H. (2019). Can we solve Ais' trust problem?. *MIT Sloan Management Review*. <https://sloanreview.mit.edu/article/can-we-solve-ais-trust-problem/>.
76. David, R. J., & Han, S.-K. (2004). A systematic assessment of the empirical support for transaction cost economics. *Strategic Management Journal*, 25(1), 39-58.
77. DeCharms, R. (1968). Personal causation. New York: AcademicPress.
78. Diakopoulos, N., & Koliska, M. (2017). Algorithmic transparency in the news media. *Digital journalism*, 5(7), 809-828.
79. Dieterle, E., Dede, C., & Walker, M. (2022). The cyclical ethical effects of using artificial intelligence in education. *AI & society*, 1-11.

80. Dilmegani, C., (2023). 9 epic Chatbot/Conversational Bot Failures. Almultiple.
<https://research.aimultiple.com/chatbot-fail/>
81. Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018, December). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 67-73).
82. Dolata, M., Feuerriegel, S., & Schwabe, G. (2022). A sociotechnical view of algorithmic fairness. *Information Systems Journal*, 32(4), 754-818.
83. Dratsch, T., Chen, X., Rezazade Mehrizi, M., Kloeckner, R., Mähringer-Kunz, A., Püsken, M., ... & Pinto dos Santos, D. (2023). Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology*, 307(4), e222176.
84. 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*, 1(2), 89-103.
85. Duguay, S. (2019). Running the numbers': Modes of microcelebrity labor in queer women's self-representation on Instagram and vine. *Social Media + Society*, 5(4), 1–11.
<https://doi.org/10.1177/2056305119894002>
86. Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... & Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57, 101994.
87. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226).
88. Eaglin, J. M. (2017). Constructing recidivism risk. *Emory LJ*, 67, 59.
89. Easton, D. (1975). A re-assessment of the concept of political support. *British journal of political science*, 5(4), 435-457.
90. Edelman, B. G., & Luca, M. (2014). Digital discrimination: The case of Airbnb. com. Harvard Business School NOM Unit Working Paper, (14-054).
91. Ethnicity and diagnosis in patients with affective disorders. *Journal of Clinical Psychiatry* 64 (7): 747–754. <https://doi.org/10.4088/jcp.v64n0702>.
92. European Commission. (2019). Ethics guidelines for trustworthy AI. Available at: <https://digital-strategy.ec.europa.eu/en/library/ethicsguidelines-trustworthy-ai> (accessed 18 July 2022).
93. Fahimnia, B., Sarkis, J., & Davarzani, H. (2015). Green supply chain management: A review and bibliometric analysis. *International Journal of Production Economics*, 162, 101-114.
94. Faris, R., Roberts, H., Etling, B., Bourassa, N., Zuckerman, E., & Benkler, Y. (2017). Partisanship, propaganda, and disinformation: Online media and the 2016 US presidential election. Berkman Klein Center Research Publication, 6.Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015, August). Certifying and removing disparate impact. In *proceedings*

- of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 259-268).
95. Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology*, 32(2), 185-193.
 96. Floridi, L., & Taddeo, M. (2016). What is data ethics?. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083), 20160360.
 97. Fritsch, S. J., Blankenheim, A., Wahl, A., Hetfeld, P., Maassen, O., Deffge, S., ... & Bickenbach, J. (2022). Attitudes and perception of artificial intelligence in healthcare: A cross-sectional survey among patients. *Digital health*, 8, 20552076221116772.
 98. Frow, P., Payne, A., Wilkinson, I. F., & Young, L. (2011). Customer management and CRM: Addressing the dark side. *Journal of Services Marketing*, 25(2), 79–89. doi:10.1108/08876041111119804
 99. Fu, R., Huang, Y., & Singh, P. V. (2020). Artificial intelligence and algorithmic bias: Source, detection, mitigation, and implications. In *Pushing the Boundaries: Frontiers in Impactful OR/OM Research* (pp. 39-63). INFORMS.
 100. Gándara, D., Anahideh, H., Ison, M. P., & Tayal, A. (2023). Inside the Black Box: Detecting and Mitigating Algorithmic Bias across Racialized Groups in College Student-Success Prediction. arXiv preprint arXiv:2301.03784.
 101. Gallagher, M. L. B. (2005). *The relationship of role strain, personal control/decision latitude, and work-related social support to the job satisfaction of distance nurse educators*. Widener University School of Nursing.
 102. Garvie, C. (2016). *The perpetual line-up: Unregulated police face recognition in America*. Georgetown Law, Center on Privacy & Technology.
 103. Gershgorn, D. (2017). AI is now so Complex its Creators can't Trust why it Makes Decisions'. <https://qz.com/1146753/ai-is-now-so-complex-its-creators-cant-trust-why-it-makes-decisions/>.
 104. Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 178(11), 1544-1547.
 105. Gillespie, Tarleton. "The relevance of algorithms." *Media technologies: Essays on communication, materiality, and society* 167.2014 (2014): 167.
 106. Glass, A., McGuinness, D. L., & Wolverton, M. (2008, January). Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on Intelligent user interfaces* (pp. 227-236).
 107. Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121-127.
 108. Gooden, S. T. (2015). *Race and social equity: A nervous area of government*. Routledge.

109. Goodman, B. W. (2016, June). Economic models of (algorithmic) discrimination. In *29th conference on neural information processing systems* (Vol. 6).
110. Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a “right to explanation”. *AI magazine*, 38(3), 50-57.
111. Green, B. (2019, December). Good” isn’t good enough. In Proceedings of the AI for Social Good workshop at NeurIPS (Vol. 17).
112. Greenhill, K. M., & Oppenheim, B. (2017). Rumor has it: The adoption of unverified information in conflict zones. *International Studies Quarterly*, 61(3), 660-676.
113. Greenwood, B., Adjerid, I., & Angst, C. M. (2017). Race and gender bias in online ratings: An origins story.
114. Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2016, December). The case for process fairness in learning: Feature selection for fair decision making. In NIPS symposium on machine learning and the law (Vol. 1, No. 2, p. 11).
115. Grimmelikhuijsen, S., & Meijer, A. (2022). Legitimacy of algorithmic decision-making: six threats and the need for a calibrated institutional response. *Perspectives on Public Management and Governance*, 5(3), 232-242.
116. Groves, L., Metcalf, J., Kennedy, A., Vecchione, B., & Strait, A. (2024). Auditing Work: Exploring the New York City algorithmic bias audit regime. *arXiv preprint arXiv:2402.08101*.
117. Gupta, D., & Krishnan, T. S. (2020). Algorithmic bias: Why bother. *California Manag. Rev*, 63(3).
118. Haas, C. (2019). The price of fairness-A framework to explore trade-offs in algorithmic fairness. *ICIS*.
119. Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California management review*, 61(4), 5-14.
120. Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., ... & Zalaudek, I. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of oncology*, 29(8), 1836-1842.
121. Hall, L. B., & Clapton, W. (2021). Programming the machine: gender, race, sexuality, AI, and the construction of credibility and deceit at the border. *Internet Policy Review*, 10(4), 1-23.
122. Hamilton, I. A. (2018). Why it’s totally unsurprising that Amazon’s recruitment AI was biased against women. *Business Insider*. Retrieved November 11 from [https:// www.businessinsider.com/amazon-ai-biased-against- women-no-surprise-sandra-wachter-2018-10](https://www.businessinsider.com/amazon-ai-biased-against-women-no-surprise-sandra-wachter-2018-10)
123. Hampton, L. M. (2021). Black feminist musings on algorithmic oppression. *arXiv preprint arXiv:2101.09869*.
124. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.
125. Hauer, T. (2019). Society caught in a labyrinth of algorithms: disputes, promises, and limitations of the new order of things. *Society*, 56, 222-230.

126. Hobson, P., & Bakker, J. (2019). How the heart attack gender gap is costing women's lives. *British Journal of Cardiac Nursing*, 14(11), 1-3.
127. Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop?. *Brain Informatics*, 3(2), 119-131.
128. Horowitz, M. C., & Kahn, L. (2023). Bending the Automation Bias Curve: A Study of Human and AI-based Decision Making in National Security Contexts. *arXiv preprint arXiv:2306.16507*.
129. Howard, A., & Borenstein, J. (2018). The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Science and engineering ethics*, 24, 1521-1536.
130. Hunter, I. (2020). *Rethinking the school: Subjectivity, bureaucracy, criticism*. Routledge.
131. Ibrahim, Y. (2010). The breastfeeding controversy and Facebook: Politicization of image, privacy and protest. *International Journal of E-Politics*, 1(2), 16–28. <https://doi.org/10.4018/jep.2010040102>
132. Israeli, A., & Ascarza, E. (2020). *Algorithmic bias in marketing*. Harvard Business School Technical Note 521–020.
133. J. Domanski, R. (2019, June). The AI pandorica: linking ethically-challenged technical outputs to prospective policy approaches. In *Proceedings of the 20th Annual International Conference on Digital Government Research* (pp. 409-416).
134. Jackson, B. A., Banks, D., Woods, D., & Dawson, J. C. (2017). Future-proofing justice: building a research agenda to address the effects of technological change on the protection of constitutional rights.
135. Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.
136. John-Mathews, J. M. (2021). Critical empirical study on black-box explanations in AI. *arXiv preprint arXiv:2109.15067*.
137. Johnson JA (2006) Technology and pragmatism: From value neutrality to value criticality, SSRN Scholarly Paper, Rochester, NY: Social Science Research Network Available at: <http://papers.ssrn.com/abstract=2154654> (accessed 24 August 2015).
138. Johnson, C. Y. (2022). Racial bias in a medical algorithm favors white patients over sicker black patients. In *Ethics of Data and Analytics* (pp. 10-12). Auerbach Publications.
139. Jones-Jang, S. M., & Park, Y. J. (2023). How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. *Journal of Computer-Mediated Communication*, 28(1), zmac029.
140. Jørgensen, A., Hovy, D., & Søgaard, A. (2016). Learning a POS tagger for AAVE-like language. In *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Proceedings of the conference*. Association for Computational Linguistics.
141. Juneja, P., & Mitra, T. (2021, May). Auditing e-commerce platforms for algorithmically curated vaccine misinformation. In *Proceedings of the 2021 chi conference on human factors in computing*

- systems (pp. 1-27).
142. Kakar, V., Franco, J., Voelz, J., & Wu, J. (2016). Effects of host race information on Airbnb listing prices in San Francisco.
 143. Kamishima, T., Akaho, S., & Sakuma, J. (2011, December). Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops* (pp. 643-650). IEEE.
 144. Katell, M., Young, M., Dailey, D., Herman, B., Guetler, V., Tam, A., ... & Krafft, P. M. (2020, January). Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 45-55).
 145. Kathayat, V. (2019). How Netflix uses AI for content creation and recommendation. *Medium* (September 28), <https://medium.com/swlh/how-netflix-uses-ai-for-content-creation-and-recommendation-c1919efc0af4>.
 146. Kay, M., Matuszek, C., & Munson, S. A. (2015, April). Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems* (pp. 3819-3828).
 147. Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018, July). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning* (pp. 2564-2572). PMLR.
 148. Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2019, January). An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 100-109).
 149. Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals*, 14(1), 366-410.
 150. Kertysova, K. (2018). Artificial intelligence and disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered. *Security and Human Rights*, 29(1-4), 55-81.
 151. Khan, A., Krishnan, S., & Dhir, A. (2021). Electronic government and corruption: Systematic literature review, framework, and agenda for future research. *Technological Forecasting and Social Change*, 167, 120737.
 152. Khurana, A., Alamzadeh, P., & Chilana, P. K. (2021, October). ChatrEx: Designing explainable chatbot interfaces for enhancing usefulness, transparency, and trust. In *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (pp. 1-11). IEEE.
 153. Kieslich, K., Keller, B., and Starke, C. (2022) Artificial intelligence ethics by design. Evaluating public perception on the importance of ethical design principles of artificial intelligence. *Big Data & Society* 9(1): 1–19. DOI: 10.1177/20539517221092956
 154. Kim, S., Lee, J., & Oh, P. (2023). Questioning AI: How Racial Identity Shapes the Perceptions of Algorithmic Bias.

155. Kiritchenko, S., & Mohammad, S. M. (2018). Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.
156. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.
157. Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, 133(1), 237-293.
158. Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, 13(3), 795-848.
159. Koene, A., Dowthwaite, L., & Seth, S. (2018, May). IEEE P7003™ standard for algorithmic bias considerations: work in progress paper. In *Proceedings of the international workshop on software fairness* (pp. 38-41).
160. Kolkman, D. (2022). The (in) credibility of algorithmic models to non-experts. *Information, Communication & Society*, 25(1), 93-109.
161. Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, 31(3), 388-409.
162. Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15), 5802-5805..
163. Kraemer F, van Overveld K, Peterson M (2011) Is there an ethics of algorithms?. *Ethics and Information Technology* 13(3): 251–260.
164. Kuang, C. (2017). Can AI be taught to explain itself. *The New York Times*, 21. <https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html>.
165. Kumar, P., Dwivedi, Y. K., & Anand, A. (2021). Responsible artificial intelligence (AI) for value formation and market performance in healthcare: The mediating role of patient's cognitive engagement. *Information Systems Frontiers*, 1-24.
166. Kumar, V., Rajan, B., Venkatesan, R., & Lecinski, J. (2019). Understanding the role of artificial intelligence in personalized engagement marketing. *California Management Review*, 61(4), 135-155.
167. Kuniavsky, M. (2010). *Smart things: ubiquitous computing user experience design*. Elsevier.
168. Kupfer, C., Prassl, R., Fleiß, J., Malin, C., Thalmann, S., & Kubicek, B. (2023). Check the box! How to deal with automation bias in AI-based personnel selection. *Frontiers in Psychology*, 14, 1118723.
169. Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in neural information processing systems*, 30.
170. Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in neural information processing systems*, 30.
171. Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management science*, 65(7), 2966-2981.

172. Latzer, M., Hollnbuchner, K., Just, N., & Saurwein, F. (2016). The economics of algorithmic selection on the Internet. In *Handbook on the Economics of the Internet* (pp. 395-425). Edward Elgar Publishing.
173. LAW, C. (1972). CIVIL RICHTS Casenote: Civil rights--restricting the use of general aptitude tests as employment criteria.(Griggs v. DUke Power Co., 401 US 424, 1971.). 3 Seton 143-158. *AJCL*, 38, 52.
174. Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380), 1094-1096.
175. Lee, M. K., Jain, A., Cha, H. J., Ojha, S., & Kusbit, D. (2019). Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1-26.
176. Lee, M., Kim, J., & Lizarondo, L. (2017). A human-centered approach to algorithmic services: Considerations for fair and motivating smart community service management that allocates donations to non-profit organizations. *Proceedings of the 2017 CHI conference on human factors in computing systems*, Denver, CO, USA
177. Lee, N. T., Resnick, P., & Barton, G. (2018). Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. Brookings Institute: Washington, DC, USA, 2.
178. Lee, N., Madotto, A., & Fung, P. (2019, August). Exploring Social Bias in Chatbots using Stereotype Knowledge. In *WNLP@ ACL* (pp. 177-180).
179. Lei, Jing, Niraj Dawar, and Zeynep Gu`rhan-Canli (2012), "Base-Rate Information in Consumer Attributions of Product-Harm Crises," *Journal of Marketing Research*, 49 (3), 336–48.
180. Lemon, K. N., & Verhoef, P. C. (2016). Understanding customer experience throughout the customer journey. *Journal of marketing*, 80(6), 69-96.
181. Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology*, 31, 611-627.
182. Liao, Q. V., Gruen, D., & Miller, S. (2020, April). Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-15).
183. Lin, C., Gao, Y., Ta, N., Li, K., & Fu, H. (2023). Trapped in the search box: An examination of algorithmic bias in search engine autocomplete predictions. *Telematics and Informatics*, 85, 102068.
184. Lopez, A., & Garza, R. (2023). Consumer bias against evaluations received by artificial intelligence: the mediation effect of lack of transparency anxiety. *Journal of Research in Interactive Marketing*.
185. Luger, E., & Sellen, A. (2016, May). " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 5286-5297).
186. Maddox, J. (2022). Micro-celebrities of information: mapping calibrated expertise and knowledge influencers among social media veterinarians. *Information, Communication & Society*, 1-27.

187. Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research*, 120, 262-273.
188. Margetts, H., and C. Dorobantu. 2019. Rethink government with AI. *Nature* 568: 163–65.
189. Martin, K. (2019a). Designing ethical algorithms. *MIS Quarterly Executive* June., 18(5), 2. <https://aisel.aisnet.org/misqe/vol18/iss2/5/>
190. Martin, F., Dwyer, T., & Martin, F. (2019b). The business of news sharing. *Sharing News Online: Commendary Cultures and Social Media News Ecologies*, 91-127.
191. Martin, K., & Waldman, A. (2022). Are algorithmic decisions legitimate? The effect of process and outcomes on perceptions of legitimacy of AI decisions. *Journal of Business Ethics*, 1-18.
192. Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.
193. McManus, I. C., Woolf, K., Harrison, D., Tiffin, P. A., Paton, L. W., Cheung, K. Y. F., & Smith, D. T. (2020). Calculated grades, predicted grades, forecasted grades and actual A-level grades: reliability, correlations and predictive validity in medical school applicants, undergraduates, and postgraduates in a time of COVID-19. *medRxiv*, 2020-06.
194. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6), 1-35.
195. Meijer, A., and M. Wessels. 2019. Predictive policing: Review of benefits and drawbacks. *International Journal of Public Administration* 42 (12): 1031–39.
196. Micelotta, E., Lounsbury, M., & Greenwood, R. (2017). Pathways of institutional change: An integrative review and research agenda. *Journal of management*, 43(6), 1885-1910.
197. Minola, T., Criaco, G., & Cassia, L. (2014). Are youth really different? New beliefs for old practices in entrepreneurship. *International Journal of Entrepreneurship and Innovation Management*, 18(2/3), 233.
198. Mitchell, M., Baker, D., Moorosi, N., Denton, E., Hutchinson, B., Hanna, A., ... & Morgenstern, J. (2020, February). Diversity and inclusion metrics in subset selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 117-123).
199. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 2053951716679679.
200. Moura, I. (2023). Encoding normative ethics: On algorithmic bias and disability. *First Monday*.+
201. Narayanan, V. K., Zane, L. J., & Kemmerer, B. (2011). The cognitive perspective in strategy: An integrative review. *Journal of Management*, 37(1), 305-351.
202. Ng, A. (2019). Machine learning yearning: Technical strategy for ai engineers in the era of deep learning. *Retrieved online at <https://www.mlyearning.org>*.
203. Noriega-Campero, A., Bakker, M. A., Garcia-Bulle, B., & Pentland, A. S. (2019, January). Active fairness in algorithmic decision making. In *Proceedings of the 2019 AAAI/ACM Conference on AI*,

Ethics, and Society (pp. 77-83).

204. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453.
205. Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology*, 45(3), 487-501.
206. OECD. (2019). Recommendation of the Council on OECD Legal Instruments Artificial Intelligence. Paris. Available at: <https://www.oecd.ai/ai-principles> (accessed 18 July 2022).
207. O'neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.
208. Osoba, O. A., & Welser IV, W. (2017). *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation.
209. Páez, A. (2021). Negligent Algorithmic Discrimination. *Law & Contemp. Probs.*, 84, 19.
210. Pansanella, V., Rossetti, G., & Milli, L. (2022). Modeling algorithmic bias: simplicial complexes and evolving network topologies. *Applied Network Science*, 7(1), 57.
211. Papakyriakopoulos, O., & Mboya, A. M. (2023). Beyond algorithmic bias: A socio-computational interrogation of the Google search by image algorithm. *Social Science Computer Review*, 41(4), 1100-1125.
212. Park, J. H., Shin, J., & Fung, P. (2018). Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.
213. Park, M., Yu, C., & Macy, M. (2023). Fighting bias with bias: How same-race endorsements reduce racial discrimination on Airbnb. *Science Advances*, 9(6), eadd2315.
214. Peters, U. (2022). Algorithmic political bias in artificial intelligence systems. *Philosophy & Technology*, 35(2), 25.
215. Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018, April). Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-12).
216. Puntoni, S., Reczek, R. W., Giesler, M., & Botti, S. (2021). Consumers and artificial intelligence: An experiential perspective. *Journal of Marketing*, 85(1), 131-151.
217. Rabassa, V., Sabri, O., & Spaletta, C. (2022). Conversational commerce: Do biased choices offered by voice assistants' technology constrain its appropriation?. *Technological Forecasting and Social Change*, 174, 121292.
218. Radlinski, F., & Craswell, N. (2017, March). A theoretical framework for conversational search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval* (pp. 117-126).
219. Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. *Ethics and information technology*, 20(1), 5-14.

220. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., ... & Wellman, M. (2019). Machine behaviour. *Nature*, 568(7753), 477-486.
221. Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, 18(1), 25-39.
222. Rawls, J. (2001). *Justice as fairness: A restatement*. Harvard University Press
223. Reyero Lobo, P., Daga, E., Alani, H., & Fernandez, M. (2023). Semantic Web technologies and bias in artificial intelligence: A systematic literature review. *Semantic Web*, 14(4), 745-770.
224. Reynolds, M., 2019. Amazon sells 'autism cure' books that suggest children drink toxic, bleach-like substances. <https://www.wired.co.uk/article/amazonautism-fake-cure-books>
225. Richardson, S. M., Petter, S., & Carter, M. (2021). Five ethical issues in the big data analytics age. *Communications of the Association for Information Systems*, (1), 18.
226. Ritter, E. M., & Brissman, I. C. (2016). Systematic development of a proctor certification examination for the Fundamentals of Laparoscopic Surgery testing program. *The American Journal of Surgery*, 211(2), 458-463.
227. Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2021). The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. *AI & society*, 36, 59-77.
228. Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., ... & Weston, J. (2020). Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.
229. Roth, Y. (2015). 'No overly suggestive photos of any kind': content management and the policing of self in gay digital communities. *Communication, Culture, & Critique*, 8(3), 414-432. <https://doi.org/10.1111/cccr.12096>
230. Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., ... & Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*.
231. Sánchez-Monedero, J., Dencik, L., & Edwards, L. (2020, January). What does it mean to solve the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 458-468). <http://arxiv.org/abs/1910.06144>
232. Schemmer, M., Kühl, N., Benz, C., & Satzger, G. (2022). On the influence of explainable AI on automation bias. *arXiv preprint arXiv:2204.08859*.
233. Schilpzand, P., De Pater, I. E., & Erez, A. (2016). Workplace incivility: A review of the literature and agenda for future research. *Journal of Organizational Behavior*, 37(Suppl 1), S57-S88.
234. Schmidt, B. (2015). Rejecting the gender binary: a vector-space operation. Ben's Bookworm Blog.
235. Schwartz, O., (2019). Untold History of AI: Algorithmic Bias Was Born in the 1980s A medical school thought a computer program would make the admissions process fairer—but it did just the opposite. *History of technology*. <https://spectrum.ieee.org/untold-history-of-ai-the-birth-of-machine-bias>
236. Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). Towards a standard for identifying and managing bias in artificial intelligence. *NIST Special Publication*, 1270, 1-77.

237. Scott, Susan V., and Wanda J. Orlikowski. 2012. "Reconfiguring Relations of Accountability: Materialization of Social Media in the Travel Sector." *Accounting, Organizations and Society* 37, no. 1: 26-40. <https://doi.org/10.1016/j.aos.2011.11.005>.
238. Serban, I., Sordoni, A., Bengio, Y., Courville, A., & Pineau, J. (2016, March). Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 30, No. 1).
239. Serban, I. V., Sankar, C., Germain, M., Zhang, S., Lin, Z., Subramanian, S., ... & Bengio, Y. (2017). A Deep Reinforcement Learning Chatbot. CoRR abs/1709.02349 (2017). *arXiv preprint arXiv:1709.02349*.
240. Shah, H. (2018). Algorithmic accountability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128), 20170362.
241. Shang, L., Lu, Z., & Li, H. (2015). Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*.
242. Shin, D. (2019). Toward fair, accountable, and transparent algorithms: Case studies on algorithm initiatives in Korea and China. *Javnost-The Public*, 26(3), 274-290.
243. Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, 98, 277-284.
244. Shin, D. (2020). User perceptions of algorithmic decisions in the personalized AI system: perceptual evaluation of fairness, accountability, transparency, and explainability. *Journal of Broadcasting & Electronic Media*, 64(4), 541-565.
245. Shin, D., Zhong, B., & Biocca, F. A. (2020). Beyond user experience: What constitutes algorithmic experiences?. *International Journal of Information Management*, 52, 102061.
246. Shin, D., Kee, K. F., & Shin, E. Y. (2022). Algorithm awareness: Why user awareness is critical for personal privacy in the adoption of algorithmic platforms?. *International Journal of Information Management*, 65, 102494.
247. Shin, D., Lim, J. S., Ahmad, N., & Ibahrine, M. (2022b). Understanding user sensemaking in fairness and transparency in algorithms: algorithmic sensemaking in over-the-top platform. *AI & SOCIETY*, 1-14.
248. Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.
249. Sigerson, L., & Cheng, C. (2018). Scales for measuring user engagement with social network sites: A systematic review of psychometric properties. *Computers in Human Behavior*, 83, 87-105
250. Silva, S., & Kenney, M. (2018). Algorithms, platforms, and ethnic bias: An integrative essay. *Phylon* (1960-), 55(1 & 2), 9-37.
251. Simoiu, C., Corbett-Davies, S., & Goel, S. (2017). The problem of infra-marginality in outcome tests for discrimination.
252. Simon-Kerr, J. (2021). Credibility in an Age of Algorithms. *Rutgers UL Rev.*, 74, 111.

253. Sloane, M., & Moss, E. (2019). AI's social sciences deficit. *Nature Machine Intelligence*, 1(8), 330-331.
254. Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.
255. Someh, I., Davern, M., Breidbach, C. F., & Shanks, G. (2019). Ethical issues in big data analytics: A stakeholder perspective. *Communications of the Association for Information Systems*, 44(1), 34. <https://doi.org/10.17705/1CAIS.04434>
256. Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., ... & Dolan, B. (2015). A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*.
257. Speicher, T., Ali, M., Venkatadri, G., Ribeiro, F. N., Arvanitakis, G., Benevenuto, F., ... & Mislove, A. (2018, January). Potential for discrimination in online targeted advertising. In *Conference on fairness, accountability and transparency* (pp. 5-19). PMLR
258. Springer, A., & Whittaker, S. (2020). Progressive disclosure: When, why, and how do users want algorithmic transparency information?. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4), 1-32.
259. Srinivasan, R., & Sarial-Abi, G. (2021). When algorithms fail: Consumers' responses to brand harm crises caused by algorithm errors. *Journal of Marketing*, 85(5), 74-91.
260. Stahl, B. (2021). From PAPA to PAPAS and beyond: Dealing with ethics in big data, AI and other emerging technologies. *Communications of the Association for Information Systems*, 49.
261. Starke, G., De Clercq, E., & Elger, B. S. (2021). Towards a pragmatist dealing with algorithmic bias in medical machine learning. *Medicine, Health Care and Philosophy*, 24, 341-349.
262. Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2022). Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*, 9(2), 20539517221115189.
263. Stevens, W. E. (2021). Blackfishing on Instagram: Influencing and the commodification of black urban aesthetics. *SocialMedia + Society*, 7(3), 1–15. <https://doi.org/10.1177/20563051211038236>
264. Strakowski, Stephen M., Paul E. Keck, Lesley M. Arnold, Jacqueline Collins, Rodgers M. Wilson, David E. Fleck, Kimberly B. Corey, Jennifer Amicone, and Victor R. Adebimpe. 2003.
265. Strawn, G. O. (2012). Scientific Research: How Many Paradigms?. *Educause Review*, 47(3), 26.
266. Stray, J. (2023). The AI Learns to Lie to Please You: Preventing Biased Feedback Loops in Machine-Assisted Intelligence Analysis. *Analytics*, 2(2), 350-358.
267. Summers, C. A., Smith, R. W., & Reczek, R. W. (2016). An audience of one: Behaviorally targeted ads as implied social labels. *Journal of Consumer Research*, 43(1), 156-178.
268. Surden, H. (2022). Values embedded in legal artificial intelligence. *IEEE Technology and Society Magazine*, 41(1), 66-74.

269. Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4), 1-13. https://grail.cs.washington.edu/projects/AudioToObama/siggraph17_obama.pdf.
270. Swaminathan, Vanitha, Karen L. Page, and Zeynep Gu`rhan-Canli (2007), "'My' Brand or 'Our' Brand: The Effects of Brand Relationship Dimensions and Self-Construal on Brand Evaluations," *Journal of Consumer Research*, 34 (2), 248–59.
271. Sweeney, L. (2013). Discrimination in online ad delivery. *Communications of the ACM*, 56(5), 44-54.
272. Taddeo, M., & Floridi, L. (2016). The debate on the moral responsibilities of online service providers. *Science and Engineering Ethics*, 22, 1575-1603.
273. Taddeo, M., McCutcheon, T., & Floridi, L. (2019). Trusting artificial intelligence in cybersecurity is a double-edged sword. *Nature Machine Intelligence*, 1(12), 557-560.
274. Thorbecke, C. (2019). New York probing Apple Card for alleged gender discrimination after viral tweet. ABC News. Retrieved february/22/2020 from <https://abcnews.go.com/US/york-probing-apple-card-alleged-gender-discrimination-viral/story?id=66910300>
275. Torraco, R. J. (2005). Writing integrative literature reviews: Guidelines and examples. *Human resource development review*, 4(3), 356-367.
276. Torralba, A., & Efros, A. A. (2011, June). Unbiased look at dataset bias. In *CVPR 2011* (pp. 1521-1528). IEEE.
277. Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2021). The ethics of algorithms: key problems and solutions. *Ethics, Governance, and Policies in Artificial Intelligence*, 97-123.
278. Turchi, T., Malizia, A., & Borsci, S. (2024). Reflecting on Algorithmic Bias with Design Fiction: the MiniCoDe Workshops. *IEEE Intelligent Systems*.
279. Turner, J. C., & Reynolds, K. J. (2011). Self-categorization theory. *Handbook of theories in social psychology*, 2(1), 399-417.
280. Vasist, P. N., & Krishnan, S. (2022). Deepfakes: an integrative review of the literature and an agenda for future research. *Communications of the Association for Information Systems*, 51(1), 14.
281. Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2), 2053951717743530.
282. Vered, M., Livni, T., Howe, P. D. L., Miller, T., & Sonenberg, L. (2023). The effects of explanations on automation bias. *Artificial Intelligence*, 103952.
283. Verma, S., & Rubin, J. (2018). Fairness definitions explained. 2018 IEEE/ACM international workshop on software fair-ness (FairWare), Gothenburg, Sweden.
284. Vimalkumar, M., Gupta, A., Sharma, D., & Dwivedi, Y. (2021). Understanding the effect that task complexity has on automation potential and opacity: Implications for algorithmic fairness. *AIS Transactions on Human-Computer Interaction*, 13(1), 104-129.
- Shin, D. (2020). How do users

- interact with algorithm recommender systems? The interaction of users, algorithms, and performance. *Computers in human behavior*, 109, 106344.
285. Vincent, J. (2016). Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge*, 24(3), 2016.
 286. Vincent, J. (2019). Google and Microsoft warn investors that bad AI could harm their brand. *The Verge*. <https://www.theverge.com/2019/2/11/18220050/google-microsoft-ai-brand-damage-investors-10-k-filing> (accessed 26 June, 2020).
 287. Vinyals, O. (2015, July). Quoc Le: A Neural Conversational Model, Deep Learning Workshop. In *32nd International Conference on Machine Learning (ICML 2015)*.
 288. Vogl, T. M., Seidelin, C., Ganesh, B., & Bright, J. (2020). Smart technology and the emergence of algorithmic bureaucracy: Artificial intelligence in UK local authorities. *Public Administration Review*, 80(6), 946-961.
 289. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31, 841.
 290. Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Proceedings of the international AAAI conference on web and social media* (Vol. 9, No. 1, pp. 454-463).
 291. Wakefield, J. (2016). Microsoft chatbot is taught to swear on Twitter. *BBC News*, 24.
 292. Walker, K. L. (2016). Surrendering information through the looking glass: Transparency, trust, and protection. *Journal of Public Policy & Marketing*, 35(1), 144-158.
 293. Waters, A., & Miiikkulainen, R. (2014). Grade: Machine learning support for graduate admissions. *Ai Magazine*, 35(1), 64-64.
 294. Weber, M., Yurochkin, M., Botros, S., & Markov, V. (2020). Black loans matter: Distributionally robust fairness for fighting subgroup discrimination. arXiv preprint arXiv:2012.01193.
 295. Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS quarterly*, xiii-xxiii.
 296. Weld, D. S., & Bansal, G. (2019). The challenge of crafting intelligible intelligence. *Communications of the ACM*, 62(6), 70-79.
 297. Whitemore, R., & Knafl, K. (2005). The integrative review: updated methodology. *Journal of advanced nursing*, 52(5), 546-553.
 298. Wilson, B., Hoffman, J., & Morgenstern, J. (2019). Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097*.<https://arxiv.org/pdf/1902.11097.pdf>.
 299. Worswick, S. (2018). Mitsuku wins Loebner Prize 2018!. *Pandorabots-blog*. <https://medium.com/pandorabots-blog/mitsuku-wins-loebner-prize-2018-3e8d98c5f2a7>.
 300. Wright, J. L., Chen, J. Y., Barnes, M. J., & Hancock, P. A. (2016). The effect of agent reasoning transparency on automation bias: An analysis of response performance. In *Virtual, Augmented and*

- Mixed Reality: 8th International Conference, VAMR 2016, Held as Part of HCI International 2016, Toronto, Canada, July 17-22, 2016. Proceedings 8* (pp. 465-477). Springer International Publishing.
301. Xie, E., Yang, Q., & Yu, S. (2021). Cooperation and Competition: Algorithmic News Recommendations in China's Digital News Landscape.
302. Xu, W. (2019). Toward human-centered AI: a perspective from human-computer interaction. *interactions*, 26(4), 42-46.
303. Yates, R. (2016, May). Data and algorithmic bias in the web. In *Proceedings of the 8th ACM Conference on Web Science* (pp. 1-1).
304. Yu, A. (2019). How Netflix Uses AI, Data Science, and Machine Learning—From A Product Perspective. *Medium* (February 27), <https://becominghuman.ai/how-netflix-uses-ai-and-machine-learning-a087614630fe>.
305. Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017, April). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web* (pp. 1171-1180).
306. Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, 41(1), 118-132.
307. Zhang, Y., Sun, S., Galley, M., Chen, Y. C., Brockett, C., Gao, X., ... & Dolan, B. (2019). Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.
308. Zhou, L., Gao, J., Li, D., & Shum, H. Y. (2020). The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1), 53-93.
309. Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4), 1060-1089.
310. Zuboff, S. (1988). *In the age of the smart machine: The future of work and power*. Basic Books, Inc.

Figures

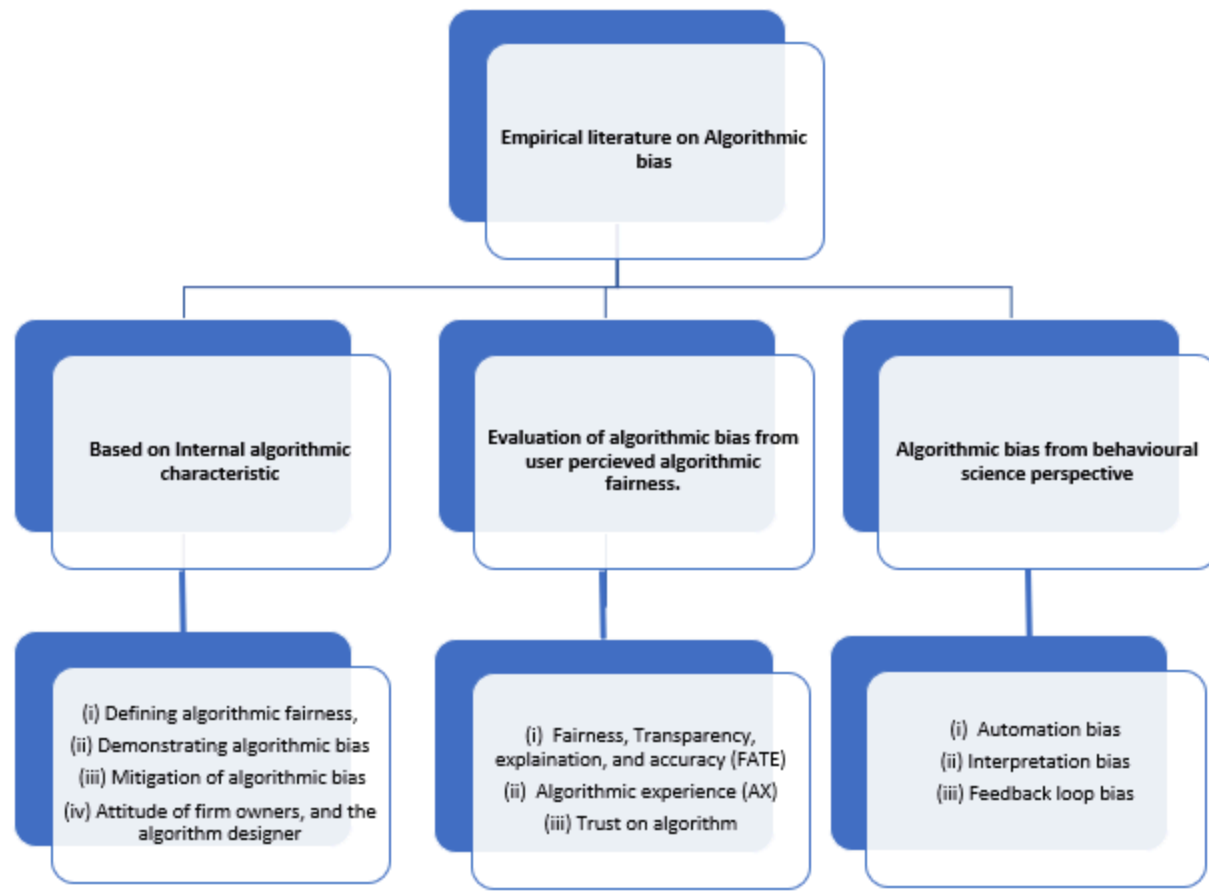


Figure 1

Classification of Empirical literature on Algorithmic Bias.

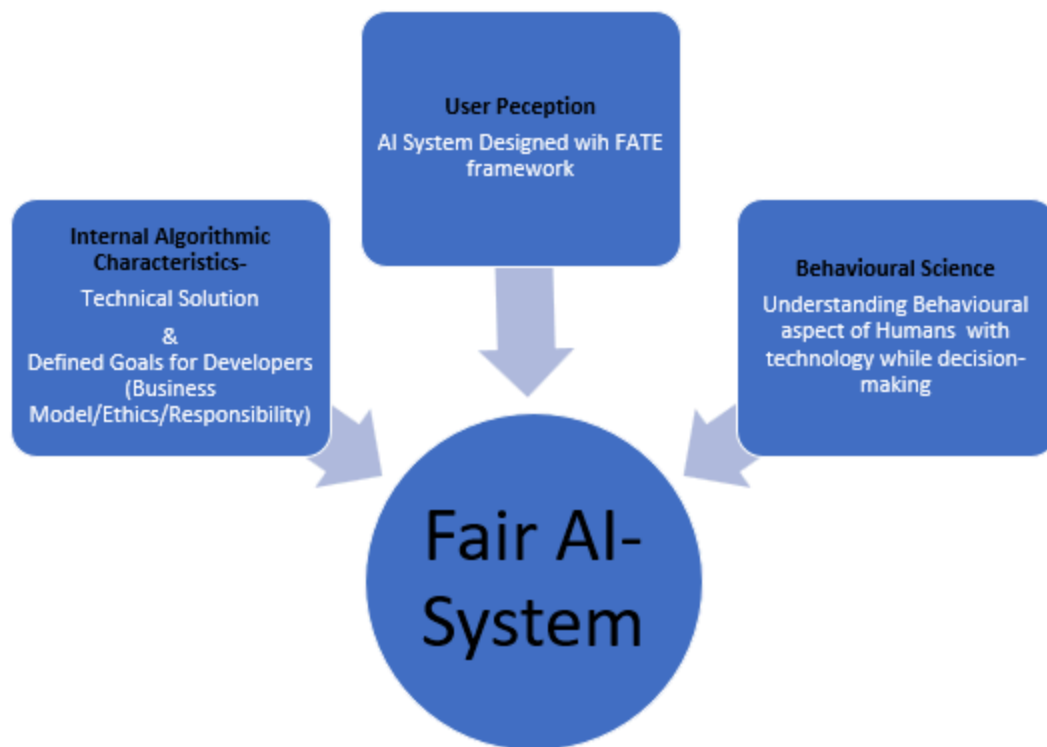


Figure 2

Integrating three aspects of literature.

Data Capturing Capability <ul style="list-style-type: none"> - Consumer Bias - Feedback loop bias 	Classification Capability <ul style="list-style-type: none"> - Non-transparency of outcome bias - Interpretation bias - Transfer of context bias - Automation bias
Delegation Capability <ul style="list-style-type: none"> - Non-transparency of outcome bias - Interpretation bias - Transfer of context bias - Automation bias 	Social Capability <ul style="list-style-type: none"> - Non-transparency of outcome bias - Interpretation bias - Transfer of context bias - Automation bias

Figure 3

Proposed Algorithmic Bias Framework (AI Capabilities and Biases)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [NewMicrosoftExcelWorksheet.xlsx](#)
- [AppendixFile.docx](#)