Mapping Scholarship on Algorithmic Bias: Conceptualization, Empirical Results, and Ethical Concerns

SEUNGAHN NAH♦[1]
University of Florida, USA

JUN LUO
University of California, Los Angeles, USA

JUNGSEOCK JOO
University of California, Los Angeles, USA

As artificial intelligence (AI) becomes more seamlessly integrated into our social life, the unfair outcomes and ethical issues associated with AI and its subtechnologies have been widely discussed in scholarly work across disciplines in recent years. This study provides an overview of the conceptualization, empirical scholarship, and ethical concerns related to algorithmic bias across diverse disciplines. In doing so, the study relies on the framework of AI-mediated communication and human-AI communication, as well as topic modeling and semantic network analysis to examine the conceptualization and major thematic areas of AI bias literature. The study reveals the complexity of the concept of algorithmic bias, which extends beyond the algorithm itself. Empirical scholarship on AI and algorithmic bias revolves around conceptualizations, human perceptions, algorithm optimization, practical applications, and ethics and policy implications. Understanding and addressing the ethical challenges require a multilevel examination from the perspectives of different stakeholders. Theoretical and practical implications are further discussed in the context of AI and algorithmic justice.

Keywords: AI, algorithmic bias, ethical issues, human-AI communication, AI-mediated communication, topic modeling, semantic network analysis

Seungahn Nah (corresponding author): snah@ufl.edu
Jun Luo: junluo0829@ucla.edu
Jungseock Joo: jjoo@comm.ucla.edu
Date submitted: 2022-11-19

[1] Seungahn Nah is Dianne Snedaker Professor in Media Trust and Research Director of the Consortium on Trust in Media and Technology at the University of Florida's College of Journalism and Communications. Jun Luo is a doctoral candidate at the UCLA's Department of Communication. Jungseock Joo is an Associate Professor at the UCLA's Department of Communication.

In the proposal that initiated the 1956 Dartmouth summer research project on artificial intelligence (AI), McCarthy, Minsky, Rochester, and Shannon (2006) outlined the fundamental components of what we now call AI: "Every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it," in which machines will be able to "use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves" (p. 12). Although their assessment in 1956 has not been fully achieved yet, subtechnologies and applications of AI have been deployed at scale—from natural language processing through computer vision to more downstream practices such as disease diagnosis, digital assistance, recommendation systems, and loan allocations. AI makes decisions about and for human beings, mediates interpersonal communication, and serves as a metric of economic growth and national security.

As AI became increasingly popular in policy decisions and public discourse, a drift toward bias, fairness, and ethical issues associated with AI technologies started to emerge in scholarly works. A large body of research has been attempting to theorize the possibilities that AI can solidify existing human cognitive and social biases and sustain unequal power relationships through decision making, interpersonal communication, and knowledge production (Bloomfield, 1988; Hancock, Naaman, & Levy, 2020; Liu, 2021; Nah, McNealy, Kim, & Joo, 2021; Noble, 2018).

Despite the diverse perspectives and approaches, ambiguity and inconsistency exist in the conceptual definition of what it means to be 'biased.' To better understand this concept, we pose a set of research questions concerning AI and algorithmic bias, focusing specifically on scholarship dealing with media and communication: How can we conceptualize and theorize algorithmic bias? What are the main areas of research on algorithmic bias? What are the ethical challenges and policy implications of algorithmic bias?

We rely on the Web of Science (WOS) database to conduct an exploratory investigation into existing definitions and conceptualization of algorithmic bias, major thematic areas in empirical studies, and the ethical challenges and policy implications. We choose WOS because the database covers a wide range of academic fields and allows users to select multiple predefined categories (i.e., WOS categories), making it suitable for an overview of interdisciplinary topics, such as AI bias in this study. Our search keywords include: (artificial intelligence OR AI OR algorithm*) AND (bias* OR fairness OR discrimination OR ethics*). To limit our search to articles published in scholarship dealing specifically with media and communication, we set the WOS categories to include any of the following four keywords: communication, media, journalism, or social. This process yields 1,517 articles. As shown in Figure 1, the number of articles on AI and algorithmic bias has tremendously increased in the past decade.

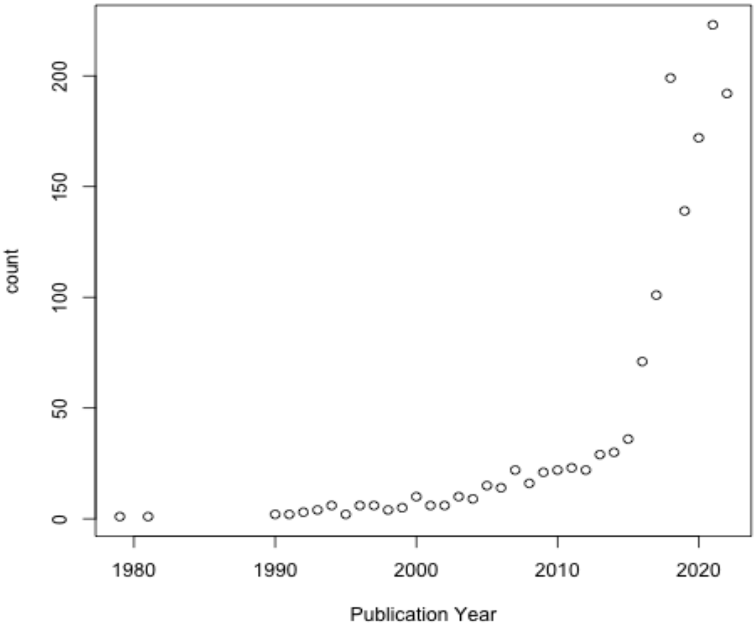

Figure 1. Number of articles on algorithmic bias by year.

To examine the main thematic areas of prior AI bias literature, we apply the mixed-method computational approach proposed by Walter and Ophir (2019). This approach combines Latent Dirichlet Allocation (LDA) topic modeling and semantic network analysis to identify news frames. The assumption is that the co-occurrence of individual topics across documents can be viewed as media frames (Walter & Ophir, 2019, pp. 249–250). The process follows three steps. First, a topic model is trained to identify the optimal number of topics in the documents and document-topic density. Second, a topic network is constructed based on co-occurrence. Lastly, network community detection algorithms are used to cluster topics (Walter & Ophir, 2019, pp. 249–253). This approach is applicable in the current study because identifying the main areas of research requires examining the common topics of interest in prior literature. Presumably, articles with similar topics tend to have identical word choice. Therefore, a topic network based on co-occurrence can capture this commonality. Thus, the topic clusters identified by network community detection algorithms can offer valuable information on the main thematic areas of extant literature.

Following their proposed process, we first train an LDA topic model on the documents using tenfold cross-validation and Gibbs sampling.[2] To better understand the materials, we then assign labels to each

[2] We ran a topic model analysis, examining various combinations of f k values (2, 5, and 10–200 with a skip of 10) and alpha values (0.01, 0.05, 0.1, 0.2, 0.5) to identify the pair of combinations that yields the lowest perplexity scores and the k value at which increasing k starts to yield diminishing returns.

topic by examining the top 50 keywords, the top 50 frequent and exclusive words, and the top 50 most representative articles for each topic, which we read in more detail. Subsequently, we construct a topic network based on co-occurrence across documents. Finally, we visualize the results of the topic clusters using the Eigen network community detection algorithm (Walter & Ophir, 2019, pp. 254–256).

In the following section, we begin with an overview of extant definitions of algorithms bias to contextualize our findings. In the second section, we present a scheme that classifies previous theorization of AI and algorithmic bias. Next, we demonstrate the thematic areas of existing scholarship using the mixed-method computational approach. Lastly, we discuss the ethical challenges and policy implications associated with algorithmic bias that we observed in prior literature.

Existing Definitions of Algorithmic Bias

Algorithmic bias is defined as "the inclination or prejudice of a decision made by an AI system which is for or against one person or group, especially in a way considered to be unfair," and these problems are "related to the gathering or processing of data that might result in prejudiced decisions on the bases of demographic features such as race, sex, and so forth" (Ntoutsi et al., 2020, p. 3). This definition highlights the unfavorable consequences based on demographic features when AI is deployed to make decisions on individuals (Ntoutsi et al., 2020; Suresh & Guttag, 2021).

Following the definition, a considerable body of prior research addresses the kinds of biases that AI systems generate in the decision-making process (Benbouzid, 2019; Brantingham, Valasik, & Mohler, 2018; Dencik, Hintz, & Carey, 2018; Obermeyer, Powers, Vogeli, & Mullainathan, 2019). For instance, in the case of health insurance allocation, if an equally sick Black patient gets less financial assistance compared to their White counterparts based on a decision made by an AI system, one can say that the algorithm is biased, and the predicted outcome is unfair based on demographic features (Obermeyer et al., 2019).

However, the definition does not consider other invisible outcomes, which are not necessarily less consequential compared to health insurance, for instance. To address this limitation, another line of research focuses on the disparity in the allocation of *symbolic* resources, such as language, image, and other audiovisual contents (Buolamwini & Gebru, 2018; Caliskan, Bryson, & Narayanan, 2017; Introna & Wood, 2004; Zhao, Wang, Yatskar, Ordonez, & Chang, 2017).

In this body of literature, algorithmic bias is discussed in relation to the harmful consequences when social groups are not equally represented in language. For instance, Blodgett, Barocas, Daumé III, and Wallach (2020) posit that representational harms occur "when a system represents some social groups in a less favorable light than others, demeans them, or fails to recognize their existence altogether" (p. 5455).

More specifically, representational harms include several subcategories: "stereotyping," "differences in system performance for different social groups," and "questionable correlations" between system behaviors and features of language associated with a particular social group (Blodgett et al., 2020,

p. 5456). Stereotyping refers to the scenario in which an AI system creates, distributes, and promotes content that contains negative generalizations about a particular social group. Associating women with certain types of words (compassionate, sensitive, etc.; Leavy, 2018) or occupations (nurse, housekeeper, secretary, etc.; Garg, Schiebinger, Jurafsky, & Zou, 2018) more often than men is one example of stereotyping.

In contrast, "differences in system performance for different social groups" and "features of language associated with a particular social group" refer to cases where predictions of language systems can vary depending on language features of different demographic groups (Blodgett et al., 2020, p. 5457). For instance, Davidson, Bhattacharya, and Weber (2019) and Sap, Card, Gabriel, Choi, and Smith (2019) find racial bias in automated hate speech detection trained on Twitter data. Tweets written in African American English are predicted as abusive at a higher rate compared to tweets in standard American English (Davidson, Bhattacharya, & Weber, 2017; Sap et al., 2019).

This definition of algorithmic bias expands the notion of "harm" denoted in the previous definition of algorithmic bias, which focuses on more visible resources. It is based on the premise that the disparity in the representation of language can also have harmful consequences. Language serves as a "frame" that activates subconscious associations between mental representations of concepts. For instance, if females are overall more likely to be associated with adjectives such as affectionate, communal, or emotional, whereas males are more likely to be linked with active, ambitious, and decisive, such implicit associations may gradually become social norms of how females (or males) are supposed to be. This difference can have a harmful impact on individual development and gender equity at large (Gaucher, Friesen, & Kay, 2011).

In addition to the disparity in the representation of language, images and videos can also be seen as "visual words" that allocate symbolic resources based on social categorization. For instance, Gutierrez (2021) proposes a typology of algorithmic gender bias in audiovisual data resulting from a mix of technological and social bias. Men are more likely to click on ads on high-paying jobs than females (referred to as interaction bias). This behavioral difference can be learned by ad recommendation systems, which, in turn, primes AI systems to target ads based on gender (referred to as presentation bias and selection bias; Gutierrez, 2021, pp. 442–443). Robot and voice assistant speech, in particular, can manifest gender bias, with young female voices more often used than male voices (Gutierrez, 2021, p. 444). In other words, machines are designed (or shaped) in a way that aligns with human biases.

Moving beyond the algorithm itself, another body of research examines how the notion and concept of algorithms influence the wider rationalities and ways of seeing the world (Beer, 2017; Liu, 2021). This type of bias is referred to as the biased perception or the belief that algorithms carry higher precision, efficiency, and objectivity, thereby having more power to shape decisions (Beer, 2017, pp. 7–9). Beer (2017) argues that the notion of an algorithm is "a vocabulary that we might see deployed to promote a certain rationality, a rationality based upon the virtues of calculation, competition, efficiency, objectivity and the need to be strategic" (Beer, 2017, p. 9), and so that an algorithm "exists not just in code but also exists in the social consciousness as a concept or term that is frequently used to stand for something" (Beer, 2017, p. 10). The consciousness that algorithms stand for accuracy and efficiency may be attributed to user-side perception and the broader discourse related to algorithms from different actors, including academics, industry practitioners, and policy makers (Eynon & Young, 2021; Liu, 2021).

The stream of research related to user-side perception frequently draws a connection to machine heuristic: the belief that machines are more objective and error-free compared to human agents (e.g., Sundar & Kim, 2019). Empirical studies also suggest that this rule of thumb guides user evaluation of AI (Gonçalves, Weber, Masullo, Silva, & Hofhuis, 2021; Wang, 2021). For instance, Gonçalves et al. (2021) examine whether the types of information being removed and the purported reason for removal have an impact on user evaluation of content moderators in the United States, the Netherlands, and Portugal. Algorithmic moderation is perceived as more just and trustworthy than human moderation, especially when no explanation is given for content removal. However, other studies have also found more nuanced results, showing that user perception of algorithms depends on the types of tasks and level of complexity (Liu & Wei, 2019; Ozanne, Bhandari, Bazarova, & DiFranzo, 2022), existing expectation of algorithms, anthropomorphism (Waddell, 2018), user prior experience with AI-powered tools (Wojcieszak et al., 2021), and trust in others (Lee, Nah, Chung, & Kim, 2020).

In terms of the broader discourse of algorithms, this line of research has found that the framing of AI technologies varies by features of information sources (Lepage-Richer & McKelvey, 2022; Shaikh & Moran, 2022). For example, Shaikh and Moran (2022) surveyed 23 U.S.-based news outlets and found that left-leaning media reported more about ethical problems of AI (privacy, surveillance, and bias and discrimination), whereas right-leaning outlets focused on the positive impacts and abuses by foreign governments (e.g., biometric data). Furthermore, AI companies can serve as important information sources for the media and thus can affect the coverage of AI technologies.

In addition to the power of the notion of the algorithm, Miceli, Posada, and Yang (2022) emphasize another dimension of algorithmic bias, that is, the power relationship involved in data design and production. The authors argue that machine-learning datasets are inherently biased due to power asymmetries among data workers, developers, and corporate forces. Reducing societal problems to fixing biases in the data or systems distracts us from the fundamental question of “who owns data and systems, who are the data workers, whose worldviews are imposed onto them, whose biases we are trying to mitigate, and what kind of power datasets perpetuate,” and most fundamentally, whether we should build AI systems in the first place (Miceli et al., 2022, p. 4). The authors, therefore, highlight the importance of having power-aware research and practices that reflect the social contexts of data design and production to understand the power asymmetries that shape the data.

Taken together, algorithmic bias is a complex construct that expands beyond the code itself and involves algorithm workers, corporate, users, media, and other stakeholders. In addition to examining different types of bias (material vs. symbolic, language vs. audiovisual) and illustrating how social bias can be reconstructed by algorithms, future research may further explore: (1) the long-term impact of disparities in symbolic resources, (2) the cognitive and psychological factors that shape users’ definition and evaluation of algorithms, (3) the cultural construction of algorithms, and (4) the power dynamics in the interrelationship among actors involved in data collection, design, and implementation processes.

How Can We Conceptualize and Theorize Algorithmic Bias?

After summarizing existing definitions of algorithmic bias, we draw from the AI communication research agenda proposed by Guzman and Lewis (2020) and Hancock et al. (2020) to schematize the conceptualizations of AI bias in three dimensions (see Figure 2).

The first level addresses micro-level decision-making processes, where machines are employed to allocate resources. This body of literature focuses on how automated algorithms can yield diverging prediction outcomes when applied on different populations and how AI designers can address such problems (Davidson et al., 2019; Obermeyer et al., 2019).

The mechanisms of micro-level AI bias can be broken down to three key components following David Marr's three-level hypothesis: the problem, training data, and algorithm architectures (Dawson, 2002). The problem refers to what specific tasks an AI algorithm is designed for. This statement specifies the rules and criteria of predictions the AI algorithm aims to make. The training data are past experiences, upon which the algorithm's future decisions are based (Heeger & Landy, 1997). Algorithm architectures denote the physical environment where the decision making is implemented under the criteria set by the problem.

Specific examples of decision-making AI include predictive policing systems, medical and healthcare AI, and more. Relevant research explains the mechanisms of decision-making algorithms, diverging outcomes based on demographic features, and the associated ethical issues. For instance, a literature survey conducted by the Mugari and Obioha (2021) summarizes the implementation of predictive policing systems and highlights the impediments: "low predictive accuracy, limited scope of crimes that can be predicted, and high cost, flawed data input, and the biased nature of some predictive software applications" are the major challenges in predictive policing systems (p. 1). In another study, Shapiro (2019) views predictive analytics as a mechanism of police reform—to rationalize patrols. The article also points out that predictive policing can be used to ameliorate human biases or capricious decision-making.

The second level is related to interpersonal communication. It centers around human-AI interaction, which includes how human-AI interactions may represent and reinforce cognitive bias and social stereotypes as discussed in the prior section (Gaucher et al., 2011; Gutierrez, 2021). AI algorithms can inherit and reflect human biases because they are designed based on the norms of human-human communication. In addition, the design of AI technologies depends on programmer objectives and user preferences. In this way, AI algorithms can replicate existing human-human communication biases. The social learning biases present in human decision-making and information-sourcing behaviors can be transmitted into algorithms through the design and training process (Kempe & Mesoudi, 2014).

Furthermore, such algorithmic bias may contribute to higher level cultural change. Presumably, the prioritized cultural traits in their designs can become normalized and affect the ways humans behave and communicate in the future if such algorithms are deployed at scale. Gmail smart replies, for instance, may encourage users to modify their ways of communicating or to normalize the behavior of being overly positive as the right way of interacting in the long term (Hancock et al., 2020; Hohenstein & Jung, 2018). This process, coupled with the belief that algorithms have higher accuracy, efficiency, and reliability, can amplify the impact of AI biases replicated from human-human communication. Such belief bias can develop into model biases when people adopt a type of behavior solely because of the authoritative nature socially assigned to machines (Kempe & Mesoudi, 2014). In the end, people may increasingly rely on the cues learned from human-AI interactions to guide their future behaviors.

Lastly, AI biases can occur at the macro level when AI serves as information gatekeepers. As indicated in mass self-communication theory, media use has become increasingly personalized as AI technologies are widely adopted in media industries, such as search engines, social media recommendation systems, and news ranking (Castells, 2007; Valkenburg, Peter, & Walther, 2016). The traditional content generation and dissemination processes can shift from two-agent interactions to one-entity intrapersonal communication as the information input of AI algorithms comes from individual media users themselves. In this context, media users may tend to select content solely based on their own needs regardless of the intent of the generator (Valkenburg et al., 2016). In the case of recommendation systems, for instance, algorithms predict user preferences based on prior browsing history and sociodemographic features to gain more user engagement. Such automation of selective exposure may result in “filter bubble” or “echo chambers,” where users are only exposed to like-minded contents (Bakshy, Messing, & Adamic, 2015; Berman & Katona, 2020; Nechushtai & Lewis, 2019; Steiner, Magin, Stark, & Geiß, 2022; Urman, Makhortykh, & Ulloa, 2022).

In sum, algorithmic bias can be understood through three levels of analysis: decision making, interpersonal communication, and information gatekeepers. These three levels can be interconnected and mutually reinforced. Biases at the decision-making level can gradually influence cues used in interpersonal communication, which can later affect what types of information is disseminated and accepted. Future research may, for instance, consider expanding the current definition of algorithmic bias and approaching the problem from these three perspectives to get a more holistic understanding.

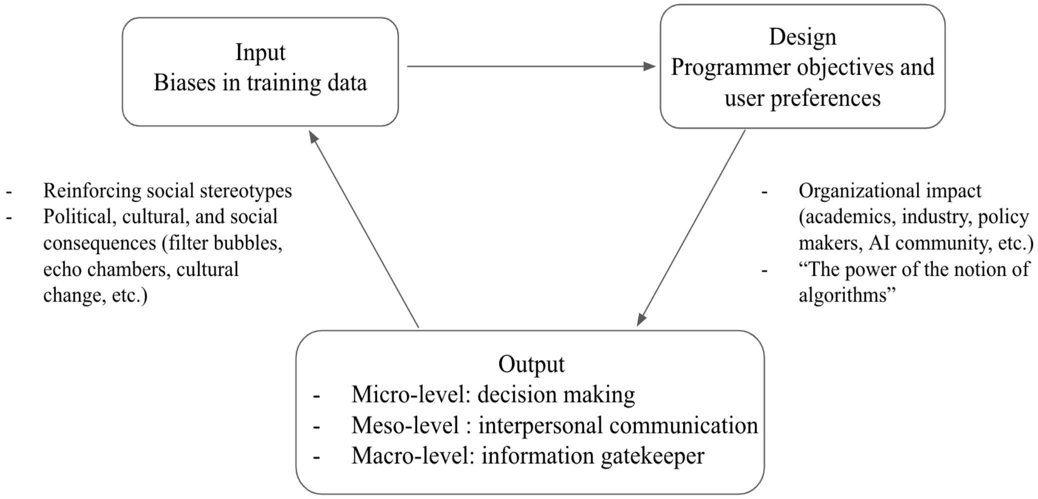
Input
Biases in training data
Design
Programmer objectives and user preferences
- Reinforcing social stereotypes
- Political, cultural, and social consequences (filter bubbles, echo chambers, cultural change, etc.)
- Organizational impact (academics, industry, policy makers, AI community, etc.)
- "The power of the notion of algorithms"
Output
- Micro-level: decision making
- Meso-level : interpersonal communication
- Macro-level: information gatekeeper

Figure 2. Algorithmic bias conceptualization.

Empirical Scholarship: What Have We Found So Far?

To provide an overview of extant empirical scholarship on AI bias, we present our results through the topic network visualized in Figure 3 where four distinct topic clusters have been identified.

The left cluster, highlighted in purple, broadly relates to the conceptualization of algorithmic bias and human perceptions of AI (Are, 2022; Moran, 2021). In addition to demographic bias manifested in decision-making AI, such as predictive policing systems, this stream of research also addresses how algorithms may bias information exposure, that is, the phenomenon of filter bubbles resulting from recommendation algorithms of search engines or social media platforms. Empirical studies have found mixed results on whether filter bubbles exist on major social media platforms and news aggregators, such as YouTube and Google News. For instance, Kaiser and Rauchfleisch (2020) map the recommendation network of 21,529 channels on YouTube. By comparing the recommendation network with that of a random network (the connection is based on random chance), they show that communities formed by YouTube recommendations have a higher homophily tendency. However, other studies find no evidence for the filter bubble impact (Haim, Graefe, & Brosius, 2018; Hosseinmardi et al., 2021). For instance, Haim et al. (2018) created four virtual agents on Google News with each mimicking the media use habits of different demographic groups. Only minor differences are found in their recommended news after a weeklong personalization process.

The adjacent cluster on the top left (in blue) shifts the focus toward the technical aspects of algorithmic bias. For instance, the topic of algorithmic optimization consists of articles explaining the black box of machine-learning algorithms and minimizing the inaccuracy and bias in the design of algorithms. Articles on other topics, such as autonomous vehicles, risk assessment, algorithmic transparency and accountability, and human opinions of algorithmic management, investigate the areas where AI and algorithms are applied along with the ethical considerations and human perceptions of algorithmic management. For example, Skeem, Scurich, and Monahan (2020) examine risk-assessment techniques in the context of criminal justice. The study tests whether risk assessments reinforce socioeconomic disparities in incarceration, which, in turn, affect judges' fairness in sentencing defendants. The results align with existing concerns, suggesting that risk-assessment tools may exacerbate sentencing disparities with affluent defendants being less likely to face incarceration compared to relatively less affluent counterparts.

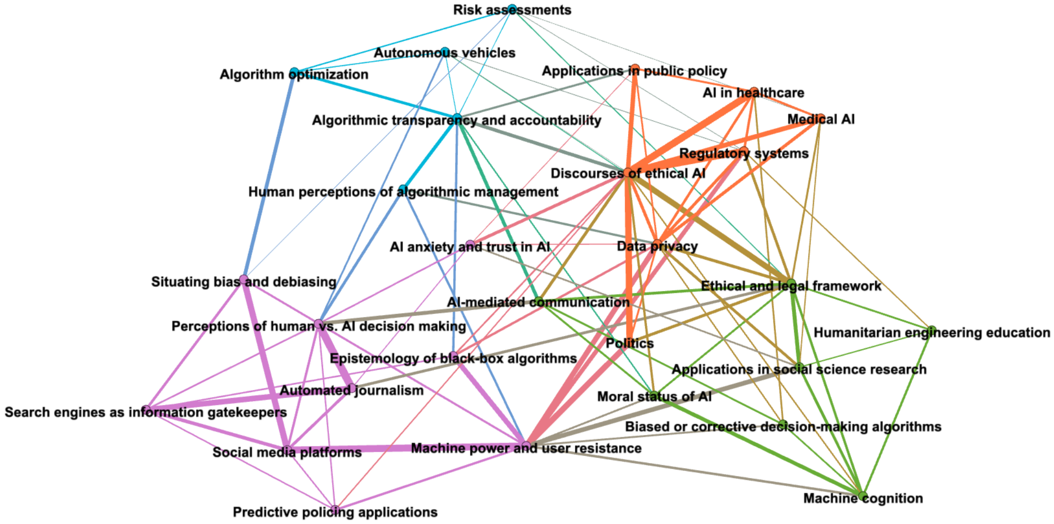
Risk assessments
Autonomous vehicles
Algorithm optimization
Applications in public policy
AI in healthcare
Medical AI
Algorithmic transparency and accountability
Regulatory systems
Discourses of ethical AI
Human perceptions of algorithmic management
AI anxiety and trust in AI
Data privacy
Situating bias and debiasing
Ethical and legal framework
AI-mediated communication
Perceptions of human vs. AI decision making
Humanitarian engineering education
Politics
Epistemology of black-box algorithms
Applications in social science research
Automated journalism
Moral status of AI
Search engines as information gatekeepers
Biased or corrective decision-making algorithms
Social media platforms
Machine power and user resistance
Machine cognition
Predictive policing applications

Figure 3. Topic network of articles on algorithmic bias.
Note. Nodes represent individual topics, edges represent co-occurrence of topics across documents, and colors represent topic communities identified by the Eigen algorithm.

In addition, the topic cluster on the top right (in orange) consists of articles delving into the social applications of AI and algorithms in the fields of healthcare, public policy, and politics. This body of work also addresses biases associated with the institutions where AI algorithms are designed or employed. For instance, by tracking Indian software engineers, Amrute (2020) examines the racialization in the tech industry where anti-immigrant violence is manifested in the division of labor. These findings resonate with prior literature using the sociological and cultural perspectives of AI bias. Human assessment of AI technologies can also be shaped by the ways different groups, such as academics, industry, and policy makers, frame AI technologies and the scientific field—the power of AI comes from both the perceived objectiveness of algorithms and the action of organizations that develop, distribute, and disseminate the technologies (Eynon & Young, 2021; Liu, 2021; Šabanović, 2014).

The final thematic area is related to AI ethics and policy implications, which we will discuss more in the next section. This body of work consists of investigations on the conceptual aspects of AI morality: whether it is reasonable to apply human morality guidelines to AI, whether AI is indeed biased, and what the possible solutions of AI bias are. For instance, Lawrence, Palacios-González, and Harris (2016) delve into the definitional puzzle of what it means to be moral for AI agents. The article explains how the morality concept that has been used to guide human-human relationships is becoming inapplicable in the case of AI. The fundamental promise of mutual respect and tolerance is based on the idea that machines possess a similar nature to human beings, which is challenged in this article. On a more practical side, Baaoum (2018) provides guidelines for attitudes, skills, and capacity building practices for humanitarian engineers. The author argues that ethics and morality rank among the most crucial attitudes perceived by respondents as essential in humanitarian engineering.

Ethical Concerns and Policy Implications

The ethical and legal discussions surrounding AI and algorithms broadly encompass the following categories. The first stream of research centers around philosophical questions of the moral and legal status of AI. Specifically, it explores whether AI possesses morality and autonomy to the extent that legal liability can be attributed to them (Bess, 2018; Kim & Kim, 2013; Serafimova, 2020). The second stream of articles unpacks the specific ethical and legal concerns associated with different types of AI technologies (Lewis, Sanders, & Carmody, 2019; McStay, 2020). The third layer focuses on who will be deemed liable for the "misconduct" of machines (Magrani, 2019; Shank, DeSanti, & Maninger, 2019). Lastly, there is an exploration of ethical guidelines, policy and legislative actions, and international and transdisciplinary collaborations that can be employed to ameliorate the issues (Feijóo et al., 2020; Kieslich, Keller, & Starke, 2022; Schaich Borg, 2021)?

With these categories in mind, we first contextualize and substantiate the ethical challenges based on literature examining AI ethical guidelines associated with each of the three dimensions of AI bias we theorized in the previous sections. An important work by Hermann (2022) categorizes the ethical principles for mass personalization systems into beneficence, nonmaleficence, justice, autonomy, and explicability using a multistakeholder perspective. By explaining potential ways that a certain ethical principle may be fulfilled or violated for different stakeholders, the author demonstrates that these principles are not independent and should be viewed from the perspectives of content senders, content receivers, and society at large (Hermann, 2022, p. 1265). Specifically, AI-powered mass personalization could be beneficent on the content sender level in terms of product satisfaction and adoption rates by optimizing content selection based on receiver preferences. However, it may be maleficent for content receivers and society as a whole because of the potential risk of selective exposure and polarization. Similarly, biases and discrimination are not limited to content receivers. Content senders, in contrast, could experience discrimination in the business domain with the case of, for instance, unequal market representation caused by the issue of filter bubbles, which violates the justice principle. Furthermore, autonomy is compromised on both content receiver and sender levels because they have delegated part of their authority to algorithms in the information filtering and dissemination process. In contrast, achieving explicability is beneficial on the side of content receivers, but it may be challenging on the sender level due to privacy and content diversity concerns. The author argues that explicability is the key principle because it serves as a prerequisite for an individual's judgments about other principles (Hermann, 2022, pp. 1266–1271). Hermann thereby proposes AI literacy as a remedy for the complex ethical challenges of AI beyond mass personalization. This principle requires all parties to have a basic understanding of data inputs and algorithmic processes, one's own capacity to decide and act, and the awareness of the potential harms of AI (Hermann, 2022, p. 1270).

After giving an overview of the ethical principles proposed by Hermann (2022), we now turn to the discussion of AI ethical issues using the three-level conceptualization we delineated in the previous section.

For micro-level decision making, ethical issues under the tenets of justice, beneficence, and nonmaleficence can arise when AI decisions are inaccurate and biased on the individual level, as well as when social inequalities are reinforced over time. For instance, the article by Benbouzid (2019) provides a comprehensive examination of predictive policing applications in the United States using the two cases of Hunchlab and PredPol. These predictive systems not only forecast crimes but also regulate police operations by producing real-time safety metrics to minimize the amount of stop-and-frisk (Benbouzid, 2019, p. 2). However, concerns that the predictions may reinforce police discrimination against minorities have arisen because the safety predictions are dependent on existing data recorded by the police. Returning to AI ethics, although predictive systems may encourage proactive policing, which optimizes police operation and increases public safety (beneficence at the practitioner level), misprediction can violate the justice principle at the individual level. Moreover, they could also reinforce the existing stereotypes associated with minority communities if the automation systems are deployed at scale in the long run (Shapiro, 2019), which harms the nonmaleficence principle from the societal perspective. In contrast, addressing the previous concerns is also challenging and requires compromise among those ethical principles, especially in the case of algorithmic protest policing. Dencik et al. (2018) argue that human interventions may exacerbate the issue of biased prediction due to the preexisting human bias of perceiving algorithms as more objective and neutral. Given that most of the software programs are often marketing-driven, it is challenging for the police to know the specificity of data inputs and the design of algorithms. Whether the software being used is applicable in a new context remains another question in this case (Dencik et al., 2018, pp. 1446–1447). These challenges highlight the importance of the explicability principle and the complexity of addressing algorithmic bias.

At the meso level, the design of interactive AI systems (e.g., gender, voice, and appearance) may reflect human preferences and stereotypes (Carpenter et al., 2009; Xu, 2019). Specifically, the language used by AI is trained on human-generated contents, which can carry social biases represented in human language (Gutierrez, 2021). Presumably, these features may be beneficial on the content sender level, enhancing user satisfaction with the product and thus increasing profit. However, continuous interactions with such agents may reinforce the existing cognitive and social biases via interpersonal communication, which can violate the nonmaleficence and justice principles at both the user and societal levels.

Lastly, the information gatekeeper role involves media organizations that employ AI-powered tools, audiences, and society at large, implying the interdependencies of different ethical principles at play in this case. For instance, in the case of algorithmic journalism, machines rely on existing data to generate news content, which is then managed and distributed by media organizations. This process in news production evokes ethical concerns on different levels (Dörr & Hollnbuchner, 2017). Before and during the content generation process, whether the data that the algorithm relies on are accurate and objective remains a question (Dörr & Hollnbuchner, 2017). Automated news production may be more efficient compared with traditional news reporting, particularly in the case of sports and finance news, but erroneous and biased reporting because inaccurate and unbalanced data can damage journalistic values such as objectivity, accuracy, fairness, and diversity. This can be harmful from the audience and society's perspective. After the content generation process, there exists a tension or compromise between individual and social sphere in terms of explicability and autonomy principles. Disclosing the data, code, and source of a news article allows extra oversight from the audience and enables them to fulfill their moral responsibility (Dörr & Hollnbuchner, 2017, p. 413), which aligns with the transparency and autonomy principle. However, the audience may selectively consume content guided by their existing perceptions about AI and professional journalists, which adds to the complexity of the issue of selective exposure.

Taken together, the ethical challenges associated with algorithmic bias are highly complex. Understanding and addressing those issues require a multilevel examination from the perspectives of different stakeholders.

With respect to specific policy implications, organizations have proposed codes of ethics and standardized guidelines with the aim to: (1) encourage diverse opinions and perspectives in data collection and model design (Leavy, 2018; Leavy, O'Sullivan, & Siapera, 2020), (2) integrate new metrics into algorithm design to avoid illegitimate discriminations (e.g., incorporating fruitless stop and frisk as negative externalities into the calculation; Benbouzid, 2019; Kasapoglu & Masso, 2021), and (3) facilitate building open datasets and implement oversight or justification mechanisms to inform the assumptions and processes of AI decision-making upfront (Benbouzid, 2019; Karppi, 2018; Orr & Davis, 2020; Williams, 2020).

However, challenges remain, particularly related to the principle of explicability. As argued by Diakopoulos and Koliska (2017), although numerous elements of AI systems could be made public, a "lack of business incentives" and the "concerns of overwhelming users with too much information" are two major obstacles to transparency in automated journalism (p. 822). In addition, existing ethical guidelines do not address the fundamental issues related to the power imbalance inherent in the AI industry. As highlighted by Miceli et al. (2022), algorithms are, in some sense, inevitably biased because of the power differentials among data workers, designers, and organizations. Therefore, future research and policy guidelines may need to pay attention to structural inequality in addition to specific techniques that aim to "fix" a single algorithm (Kerr, Barry, & Kelleher, 2020; Polack, 2020). For instance, examining (1) whether and how the market structures of AI technologies may impact the data quality and other stages of the AI lifecycle, (2) what policy actions or organizational guidelines may help encourage fair competition, (3) what practices are helpful in facilitating a fair labor market, as well as (4) providing bias- and power-aware training programs for data workers, designers, and those who operate automated systems in real-world settings (e.g., predictive policing applications).

Conclusion

This article first summarizes the existing definitions of algorithmic bias. We then theorize that AI and algorithmic bias is a multidimensional, multifaceted, and multilevel concept, which encompasses decision making at the micro level, interpersonal communication at the meso level, and mass communication at the macro level. What types of AI bias are present and how they are specifically generated depend on the level of analysis.

Moreover, using the topic modeling and semantic network analysis method, we demonstrate that existing scholarship dealing specifically with media and communication focuses on conceptualizations, human perceptions, algorithm optimization, practical applications, and ethics and policy implications. Future research may reexamine the current theorization and empirical findings under the new phenomenon of powerful conversational AI and language models.

Lastly, this article provides a review of AI ethical challenges and policy implications based on the three levels of analysis of AI or algorithmic bias and the ethical principles proposed by Hermann (2022). It reveals that understanding and addressing the ethical challenges of algorithmic bias require a thorough examination from multilevel and multistakeholder perspectives.

It is worth noting that a systematic review is not within the scope of this study. The goal of this article is to provide an initial overview of extant areas of research and provide implications for future research. Therefore, our search keywords do not include other synonyms of AI and algorithms, such as machine learning, deep learning, face recognition, or robots. Future research may, for instance, use more comprehensive keywords to conduct systematic review for research related to algorithmic bias beyond just scholarship dealing with media and communication.

References

Amrute, S. (2020). Bored techies being casually racist: Race as algorithm. *Science, Technology, & Human Values, 45*(5), 903–933. doi:10.1177/0162243920912824

Are, C. (2022). The shadowban cycle: An autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies, 22*(8), 2002–2019. doi:10.1080/14680777.2021.1928259

Baaoum, M. (2018). Humanizing engineering education: A comprehensive model for fostering humanitarian engineering education. *International Journal of Modern Education Studies, 2*(1), 1–23. Retrieved from https://dergipark.org.tr/en/pub/ijonmes/issue/38776/451409

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science, 348*(6239), 1130–1132. doi:10.1126/science.aaa1160

Beer, D. (2017). The social power of algorithms. *Information, Communication & Society, 20*(1), 1–13. doi:10.1080/1369118X.2016.1216147

Benbouzid, B. (2019). To predict and to manage. Predictive policing in the United States. *Big Data & Society, 6*(1). doi:10.1177/2053951719861703

Berman, R., & Katona, Z. (2020). Curation algorithms and filter bubbles in social networks. *Marketing Science, 39*(2), 296–316. doi:10.1287/mksc.2019.1208

Bess, M. (2018). Eight kinds of critters: A moral taxonomy for the twenty-second century. *Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine, 43*(5), 585–612. doi:10.1093/jmp/jhy018

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476). Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.485

Bloomfield, B. P. (1988). Expert systems and human knowledge: A view from the sociology of science. *AI & SOCIETY, 2*(1), 17–29. doi:10.1007/BF01891440

Brantingham, P. J., Valasik, M., & Mohler, G. O. (2018). Does predictive policing lead to biased arrests? Results from a randomized controlled trial. *Statistics and Public Policy, 5*(1), 1–6. doi:10.1080/2330443X.2018.1438940

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st conference on fairness, accountability and transparency* (pp. 77–91). Retrieved from https://proceedings.mlr.press/v81/buolamwini18a.html

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186. doi:10.1126/science.aal4230

Carpenter, J., Davis, J. M., Erwin-Stewart, N., Lee, T. R., Bransford, J. D., & Vye, N. (2009). Gender representation and humanoid robots designed for domestic use. *International Journal of Social Robotics, 1*(3), 261–265. doi:10.1007/s12369-009-0016-4

Castells, M. (2007). Communication, power and counter-power in the network society. *International Journal of Communication, 1*, 238–266.

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. In S. T. Roberts, J. Tetreault, V. Prabhakaran, & Z. Waseem (Eds.), *Proceedings of the third workshop on abusive language online* (pp. 25–35). New York, NY: Association for Computing Machinery. doi:10.48550/arXiv.1905.12516

Dawson, M. R. W. (2002). Computer modeling of cognition: Levels of analysis. In L. Nadel (Ed.), *Encyclopedia of cognitive science* (pp. 1–4). London, UK: Macmillan.

Dencik, L., Hintz, A., & Carey, Z. (2018). Prediction, pre-emption and limits to dissent: Social media and big data uses for policing protests in the United Kingdom. *New Media & Society, 20*(4), 1433–1450. doi:10.1177/1461444817697722

Diakopoulos, N., & Koliska, M. (2017). Algorithmic transparency in the news media. *Digital Journalism, 5*(7), 809–828. doi:10.1080/21670811.2016.1208053

Dörr, K. N., & Hollnbuchner, K. (2017). Ethical challenges of algorithmic journalism. *Digital Journalism, 5*(4), 404–419. doi:10.1080/21670811.2016.1167612

Eynon, R., & Young, E. (2021). Methodology, legend, and rhetoric: The constructions of AI by academia, industry, and policy groups for lifelong learning. *Science, Technology, & Human Values, 46*(1), 166–191. doi:10.1177/0162243920906475

Feijóo, C., Kwon, Y., Bauer, J. M., Bohlin, E., Howell, B., Jain, R., . . . Xia, J. (2020). Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy. *Telecommunications Policy, 44*(6), 1–14. doi:10.1016/j.telpol.2020.101988

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences, 115*(16), E3635–E3644. doi:10.1073/pnas.1720347115

Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology, 101*(1), 109–128. doi:10.1037/a0022530

Gonçalves, J., Weber, I., Masullo, G. M., Silva, M. T. d., & Hofhuis, J. (2021). Common sense or censorship: How algorithmic moderators and message type influence perceptions of online content deletion. *New Media & Society, 25*(10), 2595–2617. doi:10.1177/14614448211032310

Gutierrez, M. (2021). New feminist studies in audiovisual industries| Algorithmic gender bias and audiovisual data: A research agenda. *International Journal of Communication, 15*, 439–461.

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A Human–Machine Communication research agenda. *New Media & Society, 22*(1), 70–86. doi:10.1177/1461444819858691

Haim, M., Graefe, A., & Brosius, H. B. (2018). Burst of the filter bubble? Effects of personalization on the diversity of Google News. *Digital Journalism, 6*(3), 330–343. doi:10.1080/21670811.2017.1338145

Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication, 25*(1), 89–100. doi:10.1093/jcmc/zmz022

Heeger, D., & Landy, M. (1997). *Signal detection theory* [Teaching handout]. Department of Psychology, Stanford University. Retrieved from http://neurosci.info/courses/vision2/Coding/Heeger_2003.pdf

Hermann, E. (2022). Artificial intelligence and mass personalization of communication content—An ethical and literacy perspective. *New Media & Society, 24*(5), 1258–1277. doi:10.1177/14614448211022702

Hohenstein, J., & Jung, M. (2018). AI-supported messaging: An investigation of human-human text conversation with AI support. In R. Mandryk, M. Hancock, M. Perry, & A. Cox (Chairs), *CHI EA '18: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. LBW089:1–LBW089:6). New York, NY: Association for Computing Machinery. doi:10.1145/3170427.3188487

Hosseinmardi, H., Ghasemian, A., Clauset, A., Mobius, M., Rothschild, D. M., & Watts, D. J. (2021). Examining the consumption of radical content on YouTube. *Proceedings of the National Academy of Sciences, 118*(32), 1–8. doi:10.1073/pnas.2101967118

Introna, L., & Wood, D. (2004). Picturing algorithmic surveillance: The politics of facial recognition systems. *Surveillance & Society, 2*(2/3), 177–198. Retrieved from https://nbn-resolving.org/urn:nbn:de:0168-ssoar-200675

Kaiser, J., & Rauchfleisch, A. (2020). Birds of a feather get recommended together: Algorithmic homophily in YouTube's channel recommendations in the United States and Germany. *Social Media + Society, 6*(4), 1–15. doi:10.1177/2056305120969914

Karppi, T. (2018). "The computer said so": On the ethics, effectiveness, and cultural techniques of predictive policing. *Social Media + Society, 4*(2), 1–9. doi:10.1177/2056305118768296

Kasapoglu, T., & Masso, A. (2021). Attaining security through algorithms: Perspectives of refugees and data experts. In J. B. Wiest (Ed.), *Theorizing criminality and policing in the digital media age* (Studies in Media and Communications, Vol. 20, pp. 47–65). Bingley, UK: Emerald Publishing. doi:10.1108/S2050-206020210000020009

Kempe, M., & Mesoudi, A. (2014). Experimental and theoretical models of human cultural evolution. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*(3), 317–326. doi:10.1002/wcs.1288

Kerr, A., Barry, M., & Kelleher, J. D. (2020). Expectations of artificial intelligence and the performativity of ethics: Implications for communication governance. *Big Data & Society, 7*(1). doi:10.1177/2053951720915939

Kieslich, K., Keller, B., & 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). doi:10.1177/20539517221092956

Kim, M. S., & Kim, E. J. (2013). Humanoid robots as "the cultural other": Are we able to love our creations?. *AI & Society, 28*(3), 309–318. doi:10.1007/s00146-012-0397-z

Lawrence, D. R., Palacios-González, C., & Harris, J. (2016). Artificial intelligence: The shylock syndrome. *Cambridge Quarterly of Healthcare Ethics, 25*(2), 250–261. doi:10.1017/S0963180115000559

Leavy, S. (2018). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In E. Abraham, E. Di Nitto, & R. Mirandola (Chairs), *GE '18: Proceedings of the 1st International Workshop on Gender Equality in Software Engineering* (pp. 14–16). New York, NY: Association for Computing Machinery. doi:10.1145/3195570.3195580

Leavy, S., O'Sullivan, B., & Siapera, E. (2020, January). *Data, power and bias in artificial intelligence*. Paper presented at IJCAI 2020 AI for Social Good Workshop, Cambridge, MA. doi:10.48550/arXiv.2008.07341

Lee, S., Nah, S., Chung, D. S., & Kim, J. (2020). Predicting AI news credibility: Communicative or social capital or both? *Communication Studies, 71*(2), 1–20. doi:10.1080/10510974.2020.1779769

Lepage-Richer, T., & McKelvey, F. (2022). States of computing: On government organization and artificial intelligence in Canada. *Big Data & Society, 9*(2). doi:10.1177/20539517221123304

Lewis, S. C., Sanders, A. K., & Carmody, C. (2019). Libel by algorithm? Automated journalism and the threat of legal liability. *Journalism & Mass Communication Quarterly, 96*(1), 60–81. doi:10.1177/1077699018755983

Liu, B., & Wei, L. (2019). Machine authorship in situ. *Digital Journalism, 7*(5), 635–657. doi:10.1080/21670811.2018.1510740

Liu, Z. (2021). Sociological perspectives on artificial intelligence: A typological reading. *Sociology Compass, 15*(3), e12851. doi:10.1111/soc4.12851

Magrani, E. (2019). New perspectives on ethics and the laws of artificial intelligence. *Internet Policy Review, 8*(3), 1–19. doi:10.14763/2019.3.1420

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence. *AI Magazine, 27*(4), 12–14. doi:10.1609/aimag.v27i4.1904

McStay, A. (2020). Emotional AI, soft biometrics and the surveillance of emotional life: An unusual consensus on privacy. *Big Data & Society, 7*(1). doi:10.1177/2053951720904386

Miceli, M., Posada, J., & Yang, T. (2022). Studying up machine learning data: Why talk about bias when we mean power? *Proceedings of the ACM on Human-Computer Interaction, 6*(GROUP), 1–14. doi:10.1145/3492853

Moran, T. C. (2021). Racial technological bias and the white, feminine voice of AI VAs. *Communication and Critical/Cultural Studies, 18*(1), 19–36. doi:10.1080/14791420.2020.1820059

Mugari, I., & Obioha, E. E. (2021). Predictive policing and crime control in the United States of America and Europe: Trends in a decade of research and the future of predictive policing. *Social Sciences, 10*(6), 234, 1–14. doi:10.3390/socsci10060234

Nah, S., McNealy, J. E., Kim, J. H., & Joo, J. (2021). *Communicating artificial intelligence (AI): Theory, research, and practice*. New York, NY: Routledge.

Nechushtai, E., & Lewis, S. C. (2019). What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations. *Computers in Human Behavior, 90*, 298–307. doi:10.1016/j.chb.2018.07.043

Noble, S. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York, NY: New York University Press. doi:10.2307/j.ctt1pwt9w5

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., . . . Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), e1356. doi:10.1002/widm.1356

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. doi:10.1126/science.aax2342

Orr, W., & Davis, J. L. (2020). Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information, Communication & Society, 23*(5), 719–735. doi:10.1080/1369118X.2020.1713842

Ozanne, M., Bhandari, A., Bazarova, N. N., & DiFranzo, D. (2022). Shall AI moderators be made visible? Perception of accountability and trust in moderation systems on social media platforms. *Big Data & Society, 9*(2). doi:10.1177/20539517221115666

Polack, P. (2020). Beyond algorithmic reformism: Forward engineering the designs of algorithmic systems. *Big Data & Society, 7*(1). doi:10.1177/2053951720913064

Šabanović, S. (2014). Inventing Japan's "robotics culture": The repeated assembly of science, technology, and culture in social robotics. *Social Studies of Science, 44*(3), 342–367. doi:10.1177/0306312713509704

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1668–1678). Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-1163

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

Serafimova, S. (2020). Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement. *Humanities and Social Sciences Communications, 7*(1), 1–10. doi:10.1057/s41599-020-00614-8

Shaikh, S. J., & Moran, R. E. (2022). Recognize the bias? News media partisanship shapes the coverage of facial recognition technology in the United States. *New Media & Society*. Advance online publication. doi:10.1177/14614448221090916

Shank, D. B., DeSanti, A., & Maninger, T. (2019). When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information, Communication & Society, 22*(5), 648–663. doi:10.1080/1369118X.2019.1568515

Shapiro, A. (2019). Predictive policing for reform? Indeterminacy and intervention in big data policing. *Surveillance & Society, 17*(3/4), 456–472. doi:10.24908/ss.v17i3/4.10410

Skeem, J., Scurich, N., & Monahan, J. (2020). Impact of risk assessment on judges' fairness in sentencing relatively poor defendants. *Law and Human Behavior, 44*(1), 51–59. doi:10.1037/lhb0000360

Steiner, M., Magin, M., Stark, B., & Geiß, S. (2022). Seek and you shall find? A content analysis on the diversity of five search engines' results on political queries. *Information, Communication & Society, 25*(2), 217–241. doi:10.1080/1369118X.2020.1776367

Sundar, S. S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. In S. Brewster, G. Fitzpatrick, A. Cox, & V. Kostakos (Chairs), *CHI '19: Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–9). New York, NY: Association for Computing Machinery. doi:10.1145/3290605.3300768

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In F. Marmolejo-Cossio, R. Abebe, I. Lo, & A. A. Stoica (Chairs), *EAAMO '21: Equity and access in algorithms, mechanisms, and optimization* (pp. 1–9). New York, NY: Association for Computing Machinery. doi:10.1145/3465416.3483305

Urman, A., Makhortykh, M., & Ulloa, R. (2022). Auditing the representation of migrants in image web search results. *Humanities and Social Sciences Communications, 9*(1), 1–16. doi:10.1057/s41599-022-01144-1

Valkenburg, P. M., Peter, J., & Walther, J. B. (2016). Media effects: Theory and research. *Annual Review of Psychology, 67*(1), 315–338. doi:10.1146/annurev-psych-122414-033608

Waddell, T. F. (2018). A robot wrote this? How perceived machine authorship affects news credibility. *Digital Journalism, 6*(2), 236–255. doi:10.1080/21670811.2017.1384319

Walter, D., & Ophir, Y. (2019). News frame analysis: An inductive mixed-method computational approach. *Communication Methods and Measures, 13*(4), 248–266. doi:10.1080/19312458.2019.1639145

Wang, S. (2021). Moderating uncivil user comments by humans or machines? The effects of moderation agent on perceptions of bias and credibility in news content. *Digital Journalism, 9*(1), 64–83. doi:10.1080/21670811.2020.1851279

Williams, D. P. (2020). Fitting the description: Historical and sociotechnical elements of facial recognition and anti-black surveillance. *Journal of Responsible Innovation, 7*(sup1), 74–83. doi:10.1080/23299460.2020.1831365

Wojcieszak, M., Thakur, A., Ferreira Gonçalves, J. F., Casas, A., Menchen-Trevino, E., & Boon, M. (2021). Can AI enhance people's support for online moderation and their openness to dissimilar political views? *Journal of Computer-Mediated Communication, 26*(4), 223–243. doi:10.1093/jcmc/zmab006

Xu, K. (2019). First encounter with robot Alpha: How individual differences interact with vocal and kinetic cues in users' social responses. *New Media & Society, 21*(11–12), 2522–2547. doi:10.1177/1461444819851479

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2979–2989). New York, NY: Association for Computational Linguistics. doi:10.48550/arXiv.1707.09457