Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Psychology

ELSEVIER

Review

Algorithmic fairness as a human–technology interaction problem

Aurélie Lemmens

Algorithms shape high-stakes decisions across society. While promising efficiency, algorithms also raise fairness concerns. This article proposes that algorithmic fairness is best understood as a human–technology interaction problem rather than a purely technical challenge. Algorithms can reproduce human biases, amplify them through feedback loops, or create new forms of unfairness through objectives, proxies, and seemingly neutral variables. Yet they can make decision processes more explicit, disparities more visible, and actively mitigate discrimination. Fairness depends not only on statistical properties but also on how algorithms are designed, used and experienced by those affected by their decisions. This article therefore offers an interdisciplinary perspective that integrates insights from social justice, psychology, computer science, judgment and decision-making, and management.

Addresses
Rotterdam School of Management, Erasmus University Rotterdam, the Netherlands

Corresponding author: Lemmens, Aurélie (lemmens@rsm.nl)

Introduction

Algorithms increasingly support or replace human judgment in domains such as education, finance, healthcare, human resources, and marketing. Consider an organization that uses an algorithm to screen job applicants and decide who should be invited for an interview. Such a system promises efficiency and scalability, consistency and predictive accuracy: it can process large applicant pools, apply a consistent scoring rule, and identify profiles associated with later job performance. Yet its use also raises fundamental fairness concerns. If the algorithm is trained on historical hiring data, it may learn patterns that reflect earlier human biases, favoring applicants from certain groups, universities, or career paths that were previously overrepresented among successful hires. In doing so, it may disadvantage already marginalized groups [1].

Public debate about algorithms often oscillates between two incomplete views. One presents them as objective tools that remove human subjectivity from decision-making [2]. The other presents them as inherently discriminatory technologies that automate inequality [3]. Both views overlook a central point: algorithms are not deployed into a vacuum [4]. They are embedded within human institutions, historical inequalities, organizational incentives, and social values. These influence directly or indirectly the design and use of algorithms. This article therefore argues that algorithmic fairness is best understood not as a purely technical property of a model, but as a human–technology interaction problem [5]. Figure 1 depicts this two-way relationship: Humans shape technology through data, objectives, and governance choices, while at the same time, algorithms can help increase fairness in society by enabling bias mitigation and making disparities explicit and visible.

This article first reviews how fairness has been conceptualized in social justice and psychology and in computer science. It then describes how algorithms can both reduce or increase fairness depending on the human-technology interactions. The article concludes with implications for designing fairer algorithms.

What makes an algorithm fair?

Algorithmic fairness is difficult to define because fairness is inherently multidimensional. It is typically formalized as a comparison between groups defined by one or multiple protected attributes: characteristics such as gender, race, ethnicity, age, disability status, or religion that are socially, ethically, or legally relevant to discrimination.

Long before algorithmic fairness became a central concern, social psychology and organizational justice research showed that people evaluate fairness along multiple dimensions. Early equity theory emphasized *distributive justice*: whether outcomes are allocated in proportion to one's input or contributions [6]. Later work on *procedural justice* showed that people also care

Current Opinion in Psychology 2027, **73**:102411

This review comes from a themed issue on Humans and Technology

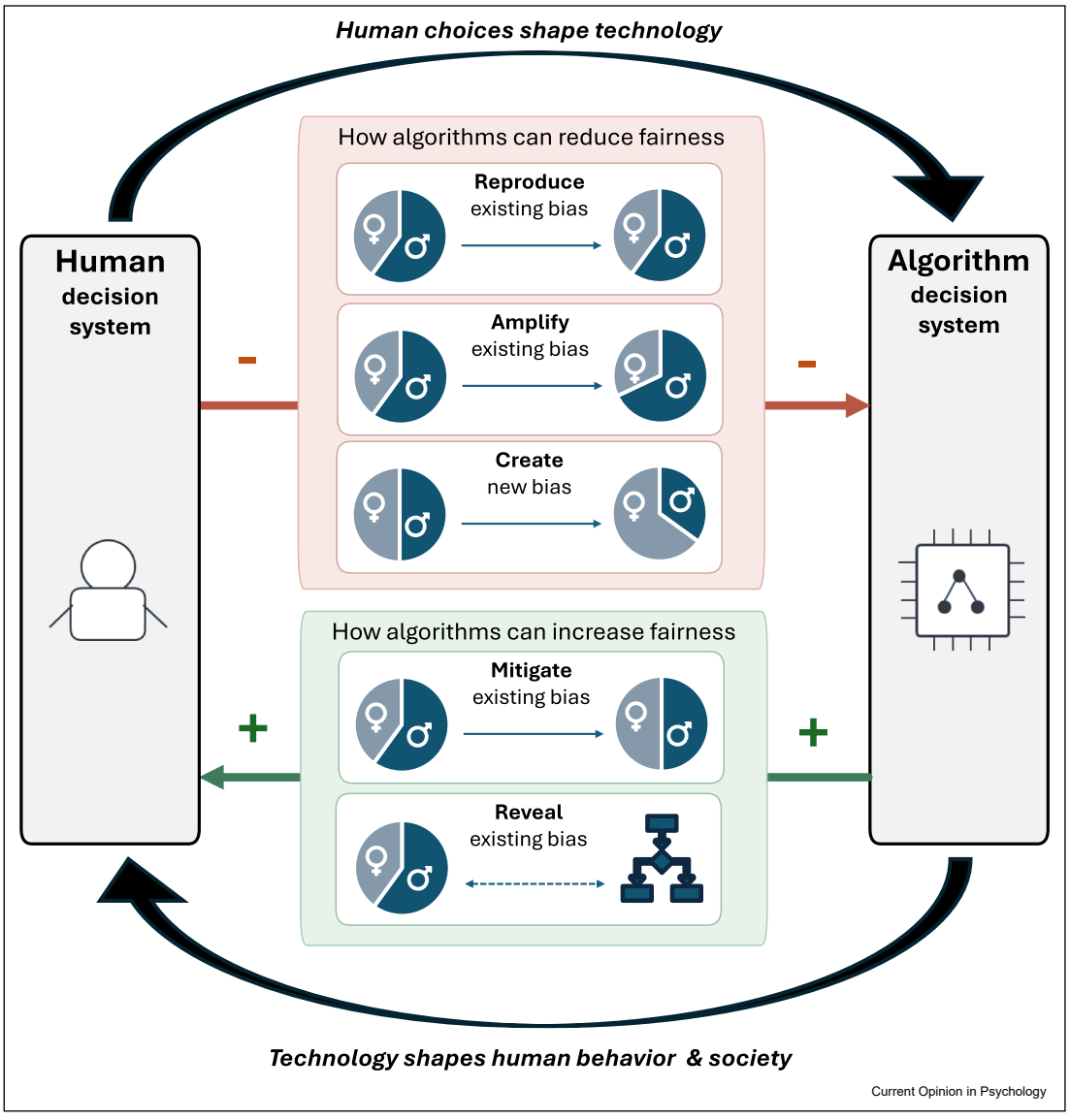
Edited by Anne-Kathrin Klesse, Ana Valenzuela, Emanuel de Bellis, and Johannes Boegershausen

For complete overview about the section, refer Humans and Technology

Available online xxx

https://doi.org/10.1016/j.copsyc.2026.102411

Figure 1



Human-Technology Interaction View on Algorithmic Fairness

Figure 1. The figure illustrates that fairness emerges from interactions between humans and algorithms. Humans shape technology through data, objectives, and governance choices. These choices lead to algorithms that reproduce, amplify, or create bias. At the same time, algorithms can help increase fairness in society by enabling bias mitigation and making disparities explicit and visible.

about how decisions are made, including whether procedures are consistent, unbiased, accurate, correctable, representative, and ethical [7]. One of the major findings of this stream of research is that the procedure to allocate outcomes has an influence on people's judgments of the fairness of a decision that is independent of outcome favorability [8]. Subsequent research further introduced the concept of *interactional justice* emphasizing that individuals are sensitive to how procedures are enacted and communicated [9]. In particular, *interpersonal justice* focuses on the extent to which individuals are treated politely and respectfully, while *informational justice* focuses, among others, on whether the decision makers provide an adequate justification: offering a causal account for a decision is associated with stronger perceptions of interactional fairness [10].

These dimensions contrast sharply with how algorithmic fairness has been defined and operationalized in computer science, where three main statistical fairness notions have emerged. First, *independence* requires that favorable outcomes, such as being invited for a job interview, occur at similar rates across groups [11]. Second, *separation* requires that error rates are similar across groups; for example,

qualified applicants should be rejected at similar rates regardless of gender or ethnicity [12]. Conceptually, separation can be linked to the merit-based notion of fairness central to distributive justice because it requires similarly qualified individuals to be treated similarly across groups. Third, *calibration* requires that the score an individual gets assigned by the algorithm has the same meaning across groups; for example, applicants who receive the same predicted job-performance score should, on average, show similar later job performance regardless of gender or ethnicity [13]. These criteria cannot be satisfied simultaneously [14]. Therefore, algorithm designers and auditors must decide which to prioritize. Unfortunately, little is known about which criterion human recipients perceive as fairest. The social justice literature has shown that people value distributive, procedural, interactional, and informational fairness differently [15], but little is known about how these concerns map onto statistical criteria [16,17]. The next section discusses the pathways through which the use of algorithms in decision-making can either reduce or increase fairness, drawing in particular on insights from justice theory and computer science.

How algorithms can reduce fairness

Algorithms can reduce fairness in three main ways: they can reproduce existing human biases, amplify those biases, or create new forms of bias that would not arise in the same way without algorithmic systems.

Reproducing existing bias

Algorithms are trained on data, often consisting of observations of past human decisions and behavior. They learn patterns in these data: how previous decisions relate to variables that characterize the individuals for whom the decisions were made. In hiring, for example, an algorithm may learn how recruiters' past screening decisions relate to applicants' education, work experience, or demographic characteristics. When those decisions reflect unequal treatment, the algorithm may interpret biased patterns as valid predictive signals thereby reproducing existing societal biases.

In the well-known Amazon recruiting case, an experimental screening tool was trained on résumés submitted over a ten-year period, most of which came from men. Consequently, it favored résumé patterns associated with male applicants while penalizing terms such as "women's" [18]. As this example illustrates, historical disparities in hiring decisions can be mechanically incorporated into an algorithm's recommendations. The problem is further compounded by the fact that applicants' educational backgrounds, even before they enter the labor market, may themselves reflect unequal opportunities arising from disparities in access to education and other resources [19].

Similar mechanisms arise in marketing: psychographic targeting algorithms trained on large text corpora may learn gendered associations in language and use them to shape which ads are shown to whom, influencing consumers' consideration sets and choices [20].

Amplifying existing bias

Algorithms do not simply reproduce human biases; they may also amplify them. This phenomenon lead to distrust and has fueled aversion toward algorithms [21]. Amplification is often caused by *feedback loops*: algorithmic decisions affect the environment from which future data are collected, thereby reinforcing the patterns the algorithm has learned [22]. In the advertising example above [20], show how targeted consumers can amplify gender bias by responding affirmatively to stereotyped ad offerings. A similar concern arises in criminal justice. If a risk assessment tool such as the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm, assigns higher risk scores to people from already over-policed groups, those scores may lead to closer monitoring and harsher decisions. This can generate more recorded contact with the justice system for those groups, reinforcing the appearance that they are higher risk [23]. Feedback loops can threaten both procedural and interactional justice. In the COMPAS case, defendants had limited ability to understand or contest the risk scores they received, especially with the feedback loops at play.

Amplification is particularly likely when systems update dynamically based on their interaction with humans, because small differences in human response can be converted into systematic disparities. This again shows why algorithmic fairness is a human–technology interaction problem: bias is not located only in the model, but can emerge from the loop between algorithmic decisions, human responses, and the new data those responses generate.

Creating new forms of bias

Algorithms may also create new forms of bias. Here, unfairness does not simply come from biased historical data or from amplifying existing disparities. It emerges from how algorithmic systems translate the decision problem into measurable objectives and proxies.

An algorithm learns patterns in the data that help optimize a given objective. In advertising, for example, this objective may be to deliver ads as cost-effectively as possible. Research has shown that this objective can result in women being underexposed to job advertisements in science, technology, engineering, and mathematics (STEM). This disparity arose not because women were explicitly excluded, but because the ad delivery system optimized exposure under cost constraints. Because women were more expensive to reach,

the optimization process generated gendered differences in access to career information [24]. Bias can therefore emerge from the objective itself, even when the advertiser does not intend to discriminate.

Bias can also emerge from the *proxies* used to measure what the algorithm is meant to predict [25]. Many decision goals are difficult to observe directly. In hiring, for example, applicants' ability to perform well on the job cannot be directly observed. In healthcare, the relevant goal may be patients' medical need, but need is complex and not always directly measured in administrative data. Designers may therefore use a proxy, a measurable variable that stands in for the true construct of interest. In the healthcare case studied by Ref. [26], an algorithm used healthcare costs as a proxy for health needs. However, because Black patients historically received less care than equally sick White patients, they generated lower costs at the same level of illness. The algorithm therefore underestimated their need for care.

Finally, new bias can also emerge when algorithms use variables that seem neutral but are closely linked to protected attributes. Removing gender, race, or ethnicity from a model does not guarantee fairness, because other variables may carry similar information. ZIP code, for example, may partly reflect ethnicity or income because residential patterns are socially stratified. In one study of online Scholastic Assessment Test (SAT) tutoring, prices varied by ZIP code, and Asian residents were nearly twice as likely as non-Asian residents to live in areas offered higher prices [27].

How algorithms can increase fairness

The human–technology interactions that make algorithms harmful can also make them valuable tools for increase fairness, either by designing fairness-aware algorithms or by enabling algorithm auditing [28].

Mitigating bias

Recent work in computer science offers opportunities for designing fairer algorithms and reducing bias. First, algorithm designers can change what the algorithm learns from, for example by rebalancing training data so that protected groups are adequately represented before the model is estimated. In experimentation and targeting, for instance, data collection can deliberately include more observations from protected groups to improve the algorithm's ability to learn meaningful response patterns for those groups [29]. Second, they can change what the algorithm is asked to optimize by integrating fairness constraints during model training, so that the system does not maximize only accuracy or profit [30]. Third, decision makers increase fairness when translating algorithmic outputs into decisions. For instance, predicted scores can be converted into binary decisions using thresholds chosen to reduce unequal error rates or unequal access across groups [12]. These developments show that fairness can be built in the system through human choices.

However, a major challenge in mitigating bias is determining which fairness definition to prioritize, as improving one dimension may weaken another. Mitigation efforts should therefore be guided by the fairness objective most appropriate to the context, which requires understanding how contextual factors shape fairness preferences. For example, independence may be more appropriate when allocating access to broadly beneficial opportunities, such as training programs, whereas separation may be more appropriate in medical diagnosis, where equally ill patients should face similar risks of error across groups.

Revealing bias

Algorithms may also support fairer decision-making because they lend themselves better to auditing than humans [31]. Human judgments vary across decision makers and contexts. For instance, recruiters may differ in how they interpret the same résumé and their decisions can be shaped by implicit judgments, influenced by contextual cues, and justified after the fact. Detecting discrimination by humans often requires carefully designed field experiments or audit studies. By contrast, once an algorithmic decision rule is specified, its behavior can be tested systematically by applying it to many cases and comparing outcomes across groups [32]. Explainable AI can further enhance this advantage by making the factors underlying algorithmic decisions more transparent, thereby helping to identify biased patterns and promote informational justice through clearer and more meaningful explanations. Recent psychological evidence even suggests that people may recognize more bias in algorithmic decisions than in their own judgments [33]. In other words, algorithms may expose biases that were already present but hidden within human judgment.

Designing fair Human–Technology systems

A central goal of this article was to demonstrate the bidirectional nature of the human–technology relationship in the context of algorithmic fairness. Consumer psychology has largely studied humans as recipients of technology, examining how they perceive artificial intelligence [34] and how recommendation systems shape human judgment and behavior [35]. This article complements that perspective by emphasizing humans as actors who also shape technology.

Past cases of algorithmic bias illustrate that scandals rarely originate from technology alone. They often arise from human decisions and institutional conditions,

including both *errors of commission*—such as selecting inappropriate objectives or data—and *errors of omission*, such as failing to examine data quality or monitor group-level outcomes [36]. Future research should therefore investigate how these errors can be reduced and how awareness of the mechanisms through which algorithmic bias emerges can be strengthened [37]. It should also clarify how responsibility should be distributed among the actors involved in designing, using and overseeing algorithms [2,38]. More systematic auditing procedures are an important part of this agenda. Regulators may help reduce errors of omission by requiring organizations to evaluate algorithms and to monitor their effects across groups. Yet meaningful audit standards cannot be developed without a clearer understanding of what affected individuals regard as fair. More research is needed on how people's attention to different aspects of fairness might shape which statistical criteria people consider most legitimate across domains such as hiring, healthcare, lending, and marketing, and why these preferences differ across contexts. It should also investigate how organizations should explain choices made in the design of algorithms, reflecting concerns about interactional justice, and how algorithmic predictions are translated into decisions, reflecting concerns about procedural justice. These questions underscore the need for interdisciplinary research. Advancing algorithmic fairness requires insights from psychology and social sciences, domain knowledge from management and other applied fields, and technical expertise from computer science, statistics, and econometrics. No single discipline can adequately address the algorithmic fairness challenge.

Funding

This research was funded by a VICI grant from the NWO (Dutch Science Foundation, File no. VI.C.231.028).

Credit author statement

Aurélie Lemmens is the solo author on this paper.

Declaration of competing interest

The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

References of particular interest have been highlighted as:

* of special interest
** of outstanding interest

1. Raghavan M, Barocas S, Kleinberg J, Levy K: **Mitigating bias in algorithmic hiring: evaluating claims and practices**. *Proc ACM Conf Fairness, Accountability, and Transparency* 2020: 469–481, https://doi.org/10.48550/arXiv.1906.09208.

2. Brayne S, Christin A: **Technologies of crime prediction: the reception of algorithms in policing and criminal courts**. *Soc Probl* 2021, **68**:608–624, https://doi.org/10.1093/socpro/spaa004.

3. Lavanchy M, Reichert P, Narayanan J, Savani K: **Applicants' fairness perceptions of algorithm-driven hiring procedures**. *J Bus Ethics* 2023, **188**:125–150, https://doi.org/10.1007/s10551-022-05320-w.

4. Castilla EJ: *AI is reinventing hiring — with the same old biases. Here's how to avoid that trap*. MIT Sloan School of Management; 2025 [cited 2026 Jun 16]. Available from: https://mitsloan.mit.edu/ideas-made-to-matter/ai-reinventing-hiring-same-old-biases-heres-how-to-avoid-trap.

5. De Cremer D, Narayanan D, Nagpal M, McGuire J, Schweitzer S: **AI fairness in action: a human-computer perspective on AI Fairness in Organizations and Society**. *Int J Hum Comput Interact* 2024, **40**:1–3, https://doi.org/10.1080/10447318.2023.2273673.

6. Adams JS: **Inequity in social exchange**. In *Advances in experimental social psychology*. Edited by Berkowitz L, New York: Academic Press; 1965.

7. Leventhal GS: **What should be done with equity theory? New approaches to the study of fairness in social relationships**. In *Social exchange: advances in theory and research*. Edited by Gergen KJ, Greenberg MS, Willis RH, Boston: Springer US; 1980.

8. Folger R: **Distributive and procedural justice: combined impact of voice and improvement on experienced inequity**. *J Pers Soc Psychol* 1977, **35**:108–119.

9. Colquitt JA: **On the dimensionality of organizational justice: a construct validation of a measure**. *J Appl Psychol* 2001, **86**: 386–400, https://doi.org/10.1037/0021-9010.86.3.386.

10. Bies RJ, Shapiro DL: **Interactional fairness judgments: the influence of causal accounts**. *Soc Justice Res* 1987, **1**: 199–218, https://doi.org/10.1007/BF01048016.

11. Kamiran F, Calders TGK: **Classifying without discriminating**. In *Proc 2nd intern Conf Computer, control and communication*. IEEE; 2009:1–6, https://doi.org/10.1109/IC4.2009.4909197.

12. Hardt M, Price E, Srebro N: **Equality of opportunity in supervised learning**. In *Advances in neural information processing systems*; 2016:3323–3331, https://doi.org/10.48550/arXiv.1610.02413.

13. Chouldechova A: **Fair prediction with disparate impact: a study of bias in recidivism prediction instruments**. *Big Data* 2017, **5**:153–163, https://doi.org/10.1089/big.2016.0047.

14. Barocas S, Hardt M, Narayanan A: *Fairness and machine learning: limitations and opportunities*. MIT Press; 2023. https://fairmlbook.org/.

15. Colquitt JA, Conlon DE, Wesson MJ, Porter COLH, Ng KY: **Justice at the millennium: a meta-analytic review of 25 years of organizational justice research**. *J Appl Psychol* 2001, **86**: 425–445, https://doi.org/10.1037/0021-9010.86.3.425.

16. Bigman YE, Wilson D, Arnestad MN, Waytz A, Gray K: **Algorithmic discrimination causes less moral outrage than human discrimination**. *J Exp Psychol Gen* 2023, **152**:4–27, https://doi.org/10.1037/xge0001250.

17. Hilliard A, Guenole N, Leutner F: **Robots are judging me: perceived fairness of algorithmic recruitment tools**. *Front Psychol* 2022, **13**, 940456, https://doi.org/10.3389/fpsyg.2022.940456.

18. Dastin J: **Insight - amazon scraps secret AI recruiting tool that showed bias against women**. *Reuters* 2018. Available from: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G/; 2018.

19. ** Goya-Tocchetto D, Kay AC, Payne BK: **Can selecting the most qualified candidate be unfair? Learning about socioeconomic advantages and disadvantages reduces the perceived fairness of meritocracy and increases support for**

socioeconomic diversity initiatives in organizations. *J Exp Psychol Gen* 2024, **153**:2962–2976, https://doi.org/10.1037/xge0001525.

20. Rathee S, Banker S, Mishra A, Mishra H: **Algorithms propagate gender bias in the marketplace—with consumers' cooperation**. *J Consum Psychol* 2023, **33**, https://doi.org/10.1002/jcpy.1351. 738–631.

21. Dietvorst BJ, Simmons JP, Massey C: **Algorithm aversion: people erroneously avoid algorithms after seeing them err**. *J Exp Psychol Gen* 2015, **144**:114–126.

22. Cowgill B, Tucker CE: *Algorithmic fairness and economics*. Columbia Business School Research Paper; 2020, https://doi.org/10.2139/ssrn.3361280.

23. Green B, Chen Y: **Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts**. *Proc ACM Hum-Comput Interact* 2021, **5**:1–33, https://doi.org/10.1145/3479562.

24. Lambrecht A, Tucker C: **Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads**. *Manag Sci* 2019, **65**:2966–2981.

25. * Zanger-Tishler M, Nyarko J, Goel S: **Risk scores, label bias, and everything but the kitchen sink**. *Sci Adv* 2024, **10**: eadi8411. https://www.science.org/doi/10.1126/sciadv.adi8411.

26. Obermeyer Z, Powers B, Vogeli C, Mullainathan S: **Dissecting racial bias in an algorithm used to manage the health of populations**. *Science* 2019, **366**:447–453. https://www.science.org/doi/10.1126/science.aax2342.

27. Vafa K, Haigh C, Leung A, Yonack N: **Price discrimination in the Princeton Review's online SAT tutoring service**. *Technol Sci* 2015. Accessed via https://techscience.org/a/2015090102.

28. * Ukanwa K: **Algorithmic bias: social science research integration into the 3-D dependable AI framework**. *Curr Opin Psychol* 2024, **58**, 101836, https://doi.org/10.1016/j.copsyc.2024.101836.

29. Shi Z, Lemmens A: *Fair active learning for targeting policies*. 2026:1–50. *Working paper*.

30. Ascarza E, Israeli A: **Eliminating unintended bias in personalized policies using bias-eliminating adapted trees**. *Proc Natl Acad Sci* 2022, **119**, e2115293119, https://doi.org/10.1073/pnas.2115293119.

31. Mullainathan S: **Biased algorithms are easier to fix than biased people**. *N Y Times* 2019, Dec 6. Accessed via, https://www.cis.upenn.edu/~mkearns/teaching/ScienceDataEthics/nyt.pdf; 2019, Dec 6.

32. Kleinberg J, Ludwig J, Mullainathan S, Sunstein CR: **Algorithms as discrimination detectors**. *Proc Natl Acad Sci* 2020, **117**: 30096–30100, https://doi.org/10.1073/pnas.1912790117.

33. * Celiktutan B, Cadario R, Morewedge CK: **People see more of their biases in algorithms**. *Proc Natl Acad Sci* 2024, **121**, e2317602121, https://doi.org/10.1073/pnas.2317602121.

34. Puntoni S, Walker Reczek R, Giesler M, Botti S: **Consumers and artificial intelligence: an experiential perspective**. *J Market* 2021, **85**:131–151.

35. Lee BC, Johar GV: **The effect of personalized content in media entertainment on engagement with the domain**. *J Consum Res* 2026, **52**:892–913, https://doi.org/10.1093/jcr/ucaf018.

36. * Chanda SS, Banerjee DN: **Omission and commission errors underlying AI failures**. *AI Soc* 2024, **39**:937–960, https://doi.org/10.1007/s00146-022-01585-x.

37. ** Hickman L, Huynh C, Gass J, Booth B, Kuruzovich J, Tay L: **Whither bias goes, I will go: an integrative, systematic review of algorithmic bias mitigation**. *J Appl Psychol* 2025, **110**: 979–1000, https://doi.org/10.1037/apl0001255.

38. Martin K: **Ethical implications and accountability of algorithms**. *J Bus Ethics* 2019, **160**:835–850, https://doi.org/10.1007/s10551-018-3921-3.

Further information on references of particular interest

19. ** Across five studies, the authors show that learning about candidates' socioeconomic advantages and disadvantages reduces the perceived fairness of merit-based selection and increases support for socioeconomic diversity initiatives. This study is important because it demonstrates that fairness judgments depend not only on qualifications and outcomes, but also on how those qualifications were produced.

25. * The authors show that including more predictors does not necessarily improve risk assessment when algorithms are trained on biased proxy outcomes. They derive conditions under which excluding variables can improve predictions of the true outcome, demonstrating why conventional "kitchen sink" modeling may amplify label bias. This study is important because it identifies a precise mechanism through which seemingly reasonable design choices can generate unfair predictions.

28. * This review argues that responsible AI research should more systematically incorporate psychological and social-scientific insights into the study of algorithmic bias. Through the lens of the 3-D Dependable AI Framework, the review explores how social science can contribute to identifying and mitigating bias at the Design, Develop, and Deploy stages of the AI life cycle.

33. * The authors find that people perceive more bias in algorithms trained on their decisions than in their own judgments, partly because of the bias blind spot. This greater recognition of bias also makes people more willing to correct decisions attributed to algorithms. The study is important because it shows that algorithms can make existing human biases more visible and potentially more open to correction.

36. * This piece investigates how AI failures occur distinguishing between omission and commission errors in the inputs to the AI system, the processing logic, and the outputs from the AI system. Based on this framework, it determines corrective action.

37. ** This systematic review develops a four-stage framework covering training-data generation, model training, testing, and deployment, and identifies sources of bias and mitigation options at each stage. It is especially important because it integrates organizational, legal, computer-science, and data-science perspectives while distinguishing mitigation methods that are both empirically effective and legal.