PREPRINT—*Manuscript under review*

The Psychology of Algorithmic Bias

David M. Amodio[1], Tessa E. S. Charlesworth[2], and William J. Brady[2]

[1]Department of Psychology, University of Amsterdam
[2]Kellogg School of Business, Northwestern University

Correspondence: david.amodio@gmail.com

Keywords: artificial intelligence, algorithm, bias, prejudice, social, cognition

Abstract

As the use of AI proliferates, so too does the risk of algorithmic bias—systematic errors in AI systems that discriminate against disadvantaged social groups. Although such biases are widely documented, their psychological foundations are poorly understood. We argue that algorithmic bias arises from human social cognition and prejudice as these processes interact with AI systems. We propose a *Human-AI Loop Model* that specifies the mechanisms through which human biases infiltrate AI systems at multiple points of human-AI interaction, from training data production to the consumption of algorithmic outputs. Through these effects, AI systems can amplify and obscure human prejudices while reinforcing existing inequities. We conclude with psychology-centered strategies for disrupting this cycle and outline implications for contemporary prejudice research.

From the use of large language models (LLMs) to facial recognition and consumer recommender algorithms, humans increasingly rely on **AI systems** (see glossary) to aid their decisions: what to read, who to date, where to shop, and who to hire, surveil, or punish [1]. Yet growing evidence reveals that operations of these systems are not neutral; rather, they frequently replicate and amplify existing patterns of **discrimination** against women, racial minorities, and other marginalized groups—a phenomenon known as **algorithmic bias** [2–6].

Highly-publicized examples of algorithmic bias include a criminal risk-assessment tool that disproportionately labeled Black defendants as high risk for recidivism [7], a hiring algorithm developed by Amazon that penalized résumés containing indicators of being female [8], and facial recognition systems used by police that were more likely to misidentify dark skin-toned individuals[9,10]. These error patterns in AI systems are not random [11]; they reflect systemic distortions that disproportionately harm socially-vulnerable populations. Although concerns about algorithmic bias have prompted both regulatory policy, such as the EU's AI Act, and computational interventions [12–14], these mitigation approaches are limited by an incomplete understanding of its fundamental source—humans—and the specific pathways through which human bias infiltrates AI.

In this article, we propose that algorithmic bias is rooted in human **social cognition** and **prejudice**, and that human biases affect nearly every stage of the AI lifecycle. That is, AI systems are conceived, trained, deployed, interpreted, and consumed by humans whose judgments are shaped by well-documented biases in perception, evaluation, and decision-making. Therefore,

any effort to explain or mitigate algorithmic bias must account for the specific mechanisms through which human psychological biases interact with AI systems.

1. A Psychological Perspective on Algorithmic Bias: The Human-AI Loop Model

From a psychological perspective, algorithmic bias is not merely a computational artifact, but a novel expression of human prejudice via AI technologies. That is, it represents the transmission of prejudice between individuals and AI systems via training data, models, and decision-support tools, which in turn generate outputs that can reinforce, legitimize, and propagate existing social inequalities [6]. Accordingly, the *Human-AI Loop* model conceptualizes algorithmic bias as emerging from dynamic interactions between human cognition and computational systems (**Fig 1**). This model identifies two critical entry points through which human prejudice infiltrates AI: in AI creation, involving the production and curation of training data, and in AI consumption, involving how human users interpret and apply AI outputs in real-world decisions. Moreover, describes a cyclical effect: human biases in the consumption of AI reinforce biased behaviors, which subsequently shape the production and selection of training data, producing self-perpetuating feedback loops of inequality and discrimination.

In what follows, we describe specific mechanisms through which human prejudices influence AI systems at points of both creation and consumption. We then outline approaches for breaking this cycle, with an emphasis on psychology-based interventions aimed at reducing bias at each stage of human-AI engagement.

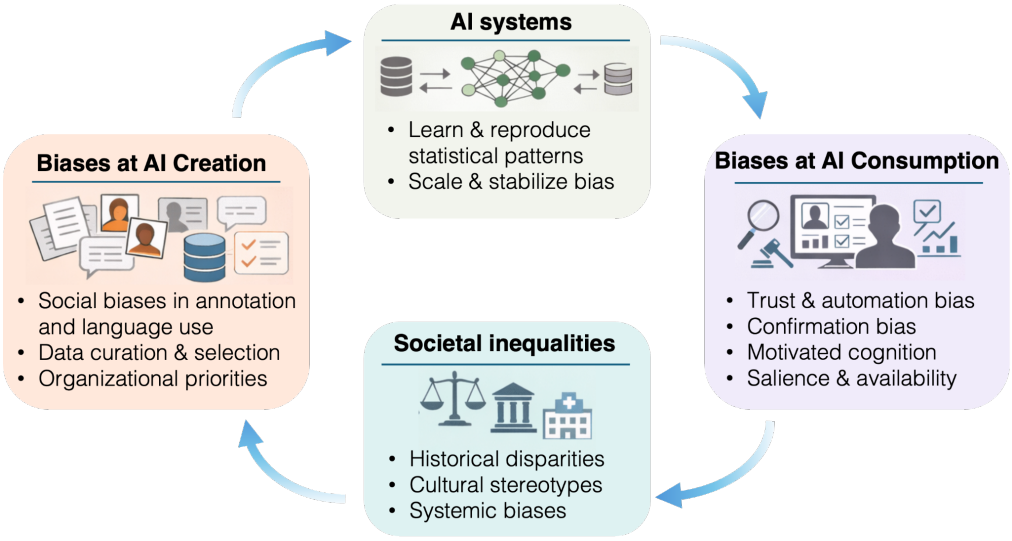


Figure 1. The "Human-AI Loop" model of algorithmic bias. Algorithmic bias is produced by human influences at the points of AI creation, through social prejudices in the production of training data and organization priorities, and AI consumption, through social and cognitive biases in the use of AI system outputs. These influences create a human-AI feedback loop through which existing societal-level inequalities are reinforced and potentially amplified via human interactions with AI systems.

2. Human Sources of Bias in AI Creation

Human prejudice first enters the AI life-cycle through (1) dataset production, (2) dataset selection, and (3) the cultures and priorities of tech organizations.

A. Production of Training Data

AI systems are only as fair as the data on which they are trained. Because most training data are generated, selected, or labeled by humans, human biases are inevitably embedded in the data used to train AI systems. Prejudices, **stereotypes**, and structural inequalities shape what data

are produced, how they are labeled, and which data are ultimately included. In this section, we describe multiple ways human biases enter AI systems via the production of training data.

i. Annotator Bias

Many supervised learning models—such as systems for automated facial recognition, content moderation, and medical diagnosis—rely on training data that are labeled by human annotators. Research in social psychology shows that the kinds of judgments annotators are asked to make—for example, judging the ethnicity of a face or the tone of a social media post—are often shaped by stereotypes, prejudices, and ideologies [15,16]. Thus, labeled AI training data are highly susceptible to these same biases [17]. Below, we describe examples of human biases in data annotation.

Face classification. Face classification systems rely on human judgments of facial attributes, such as race, gender, and emotion expression—judgments that can be influenced by stereotypes and prejudiced attitudes [16,18]. As a result, face annotation represents a critical entry point for the cultural transmission of inequality into AI systems [19]. Social psychology research has demonstrated multiple ways this can happen: **Perceptual hypodescent** is the tendency to classify a multiracial face according to its lowest-status category, functioning in part to preserve resources for dominant groups [20,21]. Perceptual hypodescent is amplified among individuals with rightwing political ideology and anti-egalitarian beliefs [22,23] and pronounced when single-category labels are required [24]. The **gendered race effect** describes the tendency for White American perceivers to classify Black faces as more masculine and East

Asian faces more feminine, reflecting American stereotypes of Black people as aggressive and Asian people as submissive [25,26]. Race and gender stereotypes also influence labels of emotion expression, too, such that male and Black and Middle Eastern faces are more readily perceived as angry or threatening, whereas female and Asian faces are more readily perceived as sad or fearful [27–29], even when expressions are objectively matched on facial muscle activity [30–32]. Finally, the **other-race effect** describes a perceiver's tendency to confuse or misidentify faces from racial outgroups [33,34]; when such errors are passed via data labeling into a face recognition training dataset, they can produce models that perform more accurately for majority-race faces.

Social Media and Content Moderation. Content moderation algorithms are used to filter harmful social media posts and to determine which posts to amplify or suppress. Creation of such algorithms relies on human annotator judgments of harm—judgments known to be influenced by a person's prejudices, stereotypes, and ideologies [35–37]. Research examining prejudice in content annotation found that Black- or Arab-authored posts were more likely to be labeled as offensive or hateful, particularly by annotators with conservative ideologies [38]. Negative stereotypes about social groups increase the tendency to label speech by those group members as hateful, and models trained on these labels reproduce these biases [39]. Similar distortions emerge when LLMs themselves are used to annotate social media [40], due to the stereotypes embedded in the text corpora used to train these labeling algorithms [41]. Thus, content annotation comprises a potent pathway through which human prejudice enters AI training data and models.

Hiring and Employment. Hiring algorithms trained on historical hiring decisions inherit the biases of human employers [42,43]. There is persistent discrimination against women and ethnic minorities in the labor market [44,45], and research shows that, even with identical résumés, women and ethnic minorities are less likely to be interviewed, evaluated positively, or hired [46,47]. AI systems trained on a company's hiring history learn to replicate these biased patterns [42]—an effect famously illustrated by Amazon's résumé-screening tool, which penalized female applicants due to the company's historical preference for male candidates [8].

Medicine and Healthcare. There are large historical racial disparities in US health care, whereby minority groups are less healthy and receive worse care [48]. Psychology research shows that physicians spend less time with minority patients, underestimate their pain, and provide lower-quality care—patterns associated with doctors' individual-level prejudice and stereotypic beliefs [49–52]. When medical AI systems are trained on data reflecting these biased decisions, they reproduce and even amplify these disparities [53–55]. For example, Obermeyer and colleagues demonstrated that under-treatment for Black American patients, represented in datasets used to train AI healthcare administration data, produced disparities of up to 46.5% between care provided to Black and White Americans [56]. Furthermore, medical diagnostic algorithms systematically underperform for women and racial minorities—underdiagnosing disease and delivering less accurate results—because training datasets disproportionately represent White male patients, which in turn reinforces existing health disparities [56–60].

Annotator Demographics and Culture. Annotators' demographic and cultural backgrounds also influence how they interpret text and images [61]. Because prejudice varies systematically

across age, gender, education, and culture [62–64], annotator pools are rarely neutral. For example, research on hate-speech classification shows that annotators' cultural stereotypes become embedded in training data and lead classifiers to reflect normative biases against marginalized groups, producing systematic errors that mirror annotators' cultural norms rather than objective definitions of harm [39]. Cultural differences further shape semantic interpretations—for example, labeling Western wedding attire as "bride" but non-Western attire as "costume"[5]—introducing subtle but consequential distortions into datasets.

Effects of Mental States, Instructions, and Incentives. Annotator judgments are further shaped by situational factors, which can amplify the expression of bias in data labeling. Fatigue, cognitive load, and low accountability—conditions under which crowd-sourced annotation may often occur—increase the influence of stereotypes on decision [65,66]. Task instructions can also induce participant bias by leading annotators to respond in ways they believe will help the work requester [67,68] or by priming particular perspectives; for example, an instruction to label based on one's personal views or knowledge could increase expressions of stereotypes [69]. Moreover, incentive structures can selectively attract annotators with particular backgrounds and motives [70]. These factors can systematically enhance bias in labeling behavior, yet they are rarely accounted for in dataset construction.

ii. Human Biases in Natural Language Data

Human language is imbued with social stereotypes and prejudices—in vocabulary, word choice, and narrative structure, as well as grammar and syntax [71–73]. Unsurprisingly, when LLMs are trained on human text corpora, they reproduce these biased patterns [74,75]. We describe

three ways in which these natural language patterns emerge in AI: through disparities in who produces text, disparities in who and what is discussed, and systematic associations between groups and attributes.

Disparities in Author Representation. Although LLMs today are largely trained on Internet-based text, training text also includes digitized books and newspapers—sources with historic representational disparities. Audit studies show that the demographics of these sources have only recently begun to approximate population-level representation [76]. For example, women were substantially underrepresented as book authors in 1970, publishing only a third as many books as men, but reached approximate parity by 2020. However, many widely used newspaper corpora (e.g., the *Wall Street Journal*) remain dominated by male authors and are written for older, high-SES, politically conservative audiences. Models trained on such data perform poorly when applied to language produced by younger or more diverse populations [77]. More broadly, minority groups tend to be represented in training data through the lens of majority-group authors who write about minorities, rather than by minority group authors themselves, which further skews minority group portrayals [78].

Language data also reflect the overrepresentation of Western, Educated, Industrialized, Rich, and Democratic (WEIRD) populations [79]. Although efforts exist to expand multilingual resources—such as translating simple dictionary approaches like LIWC into Marathi [80], or to collect multilingual texts, such as Wikipedia text from hundreds of languages [81]—non-WEIRD and lower-resource languages remain underrepresented and are served by lower-quality models [82,83]. As models are increasingly trained on their own LLM-generated text, this issue

is likely to worsen and potentially result in a "model collapse" towards just one language or culture [84].

Disparities in Content Frequency. Biases in authorship are compounded by disparities in who and what is written about in text. Men are discussed far more frequently than women, and these imbalances are amplified at intersections of gender, race, and class: for instance, only 5% of common trait descriptors are associated with Black women, whereas 59% are associated with White men [85]. As in many social domains, White, affluent men are more likely to be treated as the default category of "person" in the texts used to train LLMs [86,87].

Content disparity effects are amplified in the domain of social media, where political, moralized, emotional, and ingroup-focused content tend to be overrepresented [88]. These imbalances are further exacerbated by engagement-optimizing algorithms which, in turn, shape users' perceptions of what kinds of content are normative or acceptable to post, producing a self-reinforcing cycle of biased content generation.

Systematic Disparities in Group–Attribute Associations. Linguistic biases also arise from systematic differences in how social groups are described in the training texts. Human language disproportionately associates dominant groups with positive attributes such as competence and leadership, while linking marginalized groups to negative and subordinate attributes. Gender differences are especially salient in natural language because gender is explicitly encoded in language (e.g., pronouns such as he/she; grammatical gender in many languages) [89]. As a result, real-world gender stereotypes—such as stereotypes and societal disparities that associate women with domestic roles and men with leadership or technical roles—are

robustly reproduced in word embeddings and LLMs [12,90]. Often, these LLM stereotypes exaggerate associations beyond real-world labor distributions: for instance, while occupations like economist, pharmacist, and manager are now ~50% women (based on U.S. Bureau of Labor Statistics), they remain male-stereotyped due to default associations between men and work [90].

Although much research has studied gender bias in language, comparable patterns appear for race, immigration status, social class, and health-related stigmas, with marginalized groups more strongly associated with poverty, criminality, animalistic language, or negative affect [41,85,91,92]. Moreover, while the specific attributes linked to these groups shift over time, the negative valence attached to lower-power groups remains strikingly stable across historical periods [93]. These findings demonstrate that biased group–attribute associations are a pervasive feature of human language, rooted in human social cognition and societal structures, and consequently a persistent source of bias in language-based AI systems.

iii. Proxy Effects

Explicit efforts to avoid social biases, such as through the exclusion of social category information (e.g., on race, gender, or age), can fail due to a hidden correlation between social categories and other aspects of the data—that is, a **proxy effect** [94]. For example, zip code, occupation, education, or consumer behavior variables often serve as proxies for race, gender, or socioeconomic status because they vary systematically as a function of these demographics [95]. Recidivism tools such as COMPAS—found to predict a higher likelihood of recidivism for Black than White Americans despite equal past offenses—illustrate this dynamic: although race

was omitted from its training data, geographic variables correlated with race reproduced racial disparities in risk scores [96]. Proxy effects are particularly insidious because they obscure the presence of bias while preserving its consequences (**Box 1**).

B. Selection of Data for Model Training

The choice of dataset to use for model training is itself a human decision, guided by technical objectives, availability, convenience, and cost, and also by developers' awareness of, and concern for, the social consequences of their systems. Even when developers do not intend to introduce bias, dataset selection can systematically misrepresent or exclude vulnerable populations, which in turn can produce models that disproportionately harm those groups. In this section, we describe three contexts in which dataset choice introduces human bias into an AI system.

Selection of image training data. The impact of training dataset choice on face classification systems has been well documented. Large, widely used face image datasets—often scraped from U.S.-based websites and social media—are heavily skewed toward White faces, and models trained on such datasets perform worse for individuals with darker skin tones, especially women of color [97]. When such datasets are selected to train face classification AI systems, these systems perform less accurately in classifying and identifying minority-group faces [98]. By contrast, face classifiers trained on demographically balanced datasets, such as DiveFace [99] and FairFace [97], show substantial improvements in accuracy across race, gender, and age and reductions in performance disparities.

Selection of Text Corpora. As discussed in the previous section, LLMs are trained on digitized text drawn from multiple sources, and the composition of these corpora has profound implications for model behavior. In recent years, LLMs are increasingly trained on text from blogs, forums, and social media—media outlets accessible to a wider range of social groups. While these sources add demographic diversity, they introduce their own biases; for example, social media content is disproportionately political, moralized, emotional, and polarized [100,101]. When such datasets are selected to train an AI system, the resulting models are likely to overestimate the prevalence, normativity, or acceptability of extreme or divisive content such as overt racism and sexism.

Cohort effects. Training data selection is also shaped by *when* data were produced. Language, norms, and social attitudes change over time, and older texts can express stronger social stereotypes than more recent ones [41,77,102]. When decades-old news articles or books are selected to train LLMs, these LLMs inherit not only outdated vocabulary and grammar but also historically entrenched prejudices. These cohort effects can be exacerbated by the pursuit of scale: If data are included from older sources to increase a dataset's size, the increased size could paradoxically intensify social bias.

C. Tech Culture & Organizational Effects

Algorithmic bias is also shaped by the organizational contexts in which AI systems are developed. When AI tools are created within technology firms whose goals, incentive

structures, and cultures are misaligned with principles of fairness or equity, these organizational factors constitute an additional source of human bias in AI systems.

Lack of diversity in tech. Technology companies—particularly in the United States—remain demographically homogeneous, with leadership positions disproportionately occupied by White and Asian men [103]. Women, Black, Latino, disabled, and LGBTQ+ individuals are underrepresented, particularly in decision-making roles, and many tech workers come from relatively high socioeconomic backgrounds. Research in organizational psychology shows that such homogeneity produces predictable blind spots: male-dominated organizations are less likely to recognize harms affecting women, and White-majority organizations are more likely to overlook or discount harms to racial minorities [44,104,105]. Highly publicized failures, like the so-called 'racist soap dispensers' that failed to detect dark skin [106], illustrate these risks. Such blind spots influence which concerns are raised, which harms are anticipated, and which design decisions are prioritized or dismissed.

Market Pressures and Power Dynamics. Organizational culture interacts with market pressures to further exacerbate bias toward historically disadvantaged groups. AI systems are often optimized for engagement or efficiency in ways that do not always align with fairness or positive social impact [88]. In the absence of regulatory constraints, bias mitigation may be deprioritized—particularly when harms affect groups with limited economic or political power.

Organizational power dynamics further shape algorithmic bias through relationships with annotators, contractors, and users. Annotators—often low-paid and geographically distant—have limited power compared with tech companies, and research shows that employees who

have less power or more vulnerable work status are less likely to raise concerns about the ethics of an organization or a task [107]. The emergence of AI auditing as a professional practice reflects growing recognition that organizational structures themselves are a critical context for understanding and mitigating algorithmic bias [108].

D. Multimodal Compounding of Human Bias

So far, we have described distinct human sources of bias and their independent effects on AI systems. However, many contemporary AI systems are multimodal, integrating information from text, images, numerical data, and human-generated judgments [109]. In such systems, biases introduced at different stages and in different modalities do not merely coexist—they compound. For example, a medical decision-support system may combine radiological images, clinical notes, prior diagnoses, and demographic or environmental risk factors. If each input includes human bias—such as underrepresentation of minority patients in imaging datasets, biased language in clinicians' notes, and historically unequal access to care encoded in medical records—the resulting model may amplify disparities more strongly than any single data source [110]. Because these biases originate from distinct yet correlated human decisions, their effects can interact multiplicatively rather than additively.

Multimodal compounding also obscures sources of bias, making discrimination harder to detect and correct. Errors may appear to arise from complex model interactions rather than from identifiable human inputs, diffusing accountability across data types and organizational roles and contributing further to **AI bias laundering** (**Box 2**). Understanding how human biases

accumulate across modalities will be essential for diagnosing algorithmic bias and designing effective interventions.

3. Human Sources of Bias in AI Consumption

The second major entry point for human sources of algorithmic bias is in their downstream effects: in how AI systems are perceived, interpreted, and acted upon by human decision-makers whose judgments are shaped by well-documented biases in social cognition, motivation, and reasoning. These psychological processes systematically influence whether algorithmic outputs are trusted and implemented in ways that reinforce the very biases already encoded in AI systems.

Overreliance and Trust in AI. A key determinant of whether a biased model translates into real-world discrimination is the extent to which human users trust and accept its output. When AI systems are perceived as fair, objective, or technically superior, users are more likely to rely on them uncritically, allowing biased outputs to go unchecked and enacted in real-world decisions [111].

Trust in AI varies across individuals: Users with less knowledge or experience with search engines and AI systems tend to be more trusting of algorithmic outputs [112,113]. Trust varies by age, too, with children and older adults exhibiting greater trust in internet search results than younger adults [114], and by gender, with men trusting AI more than women [115]. Greater trust in AI, in turn, predicts stronger intentions to rely on AI in decision-making [116].

The risk posed by trust is compounded by AI's *veneer of objectivity*—the tendency to perceive mathematical or computational outputs as inherently impartial [117]. However, this effect depends on context: users are more skeptical of AI outputs when they mimic uniquely human capacities such as moral reasoning [118,119].

An additional concern is that differences in trust between sociodemographic groups (e.g., women, older people) lead to systematic differences in the adoption or skilled utilization of AI tools [120], which could disadvantage certain groups. This pattern was observed in a study in China, where wage disparities between men and women were in part attributed to a gender gap in Internet usage [121].

Motivated Social Cognition and Prejudice. Human interactions with AI can be further shaped by motivated social cognition—the tendency to process information in ways that serve one's personal or group-based goals [122,123]. Research shows that, indeed, trust in AI is influenced by one's goals and beliefs [111,124,125], suggesting that AI outputs matching an individual's existing views are more likely to be accepted, trusted, and applied in decision-making [126]. These tendencies reflect confirmation bias: the tendency to search for, prioritize, and accept information that supports preexisting beliefs or expectations [94]. In practice, this means that even identical algorithmic outputs can have different downstream effects depending on the user's goals and attitudes.

Availability and Anchoring Heuristics. Even in the absence of social motives, limitations of human cognition can produce decision biases. For example, human judgments tend to rely on information that is salient or readily accessible, especially under uncertainty—an effect known

as the availability heuristic [127]. Moreover, initial information tends to anchor one's judgment, such that it weighs more heavily than subsequent information in a decision process. These findings suggest that, in the context of AI, a decision maker would over-weigh information presented first or most prominently, such as top-ranked search results or the initial output of a generative model [94,128]. A study of gender bias in internet search outputs demonstrated this effect: the prevalence of men relative to women in search results for specific professions strongly influenced users' beliefs and decisions regarding the suitability of men (vs. women) for those jobs [6].

4. Closing the Loop: A Cycle of Bias Between Humans and AI

The Human-AI Loop Model emphasizes the cyclical nature of human and computational contributions to algorithmic bias. Humans and AI systems are not independent sources of bias, but interact dynamically to reinforce one another over time: human prejudices are embedded in training data, biased data shape model outputs, model outputs in turn shape human beliefs and behavior, which generates new data that feed into subsequent models [6,101]. Over time, this feedback loop can entrench and accelerate inequality.

Unlike prior 'human-in-the-loop' models, which aim to engage human oversight to improve fairness and accuracy [129], our analysis describes human contributions to algorithmic bias. That is, the Human-AI Loop model describes specific mechanisms through which human bias, stemming from prejudices as well as general cognitive tendencies, can introduce or amplify bias in AI systems (**Table 1**). By emphasizing the many ways in which human bias infiltrates AI, at

multiple stages of human-AI interaction, this model expands and clarifies our understanding of how algorithmic bias is formed and expressed.

Table 1. Human Sources of Bias Across the AI Use Cycle and Intervention Targets

AI Use Stage	Human Source of Bias	Core Psychological Mechanisms	Illustrative Intervention Strategies
AI Creation: Data Productions & Curation	Biased annotation	Stereotyping; implicit prejudice; heuristic judgment	Diverse annotator pools; bias-aware instructions; training with feedback; preserve annotator disagreement
	Homogeneous datasets	Availability bias; limited perspective-taking	Match data to intended use; ensure adequate or equal group representation; dataset documentation
	Historical bias in records	Institutional discrimination reflected in outcomes	Reweight or exclude biased labels; supplement with counterfactual data
	Linguistic & cultural bias	WEIRD overrepresentation; group–attribute associations	Expand non-WEIRD and multilingual corpora; report corpus composition
	Situational pressures on annotators	Fatigue; low accountability	Improved labor conditions; workload limits; quality monitoring
Organizational Context	Lack of diversity in tech organizations	Social identity norms; organizational blind spots	Inclusive hiring and leadership; participatory design practices
	Market-driven incentives	Motivated reasoning; profit optimization	Regulation; mandated impact assessments; fairness benchmarks
	Power asymmetries	Moral distancing; reduced accountability	Fair labor standards; transparency; independent AI auditing
AI Consumption: Human Use & Interpretation	Algorithmic overtrust	Automation bias; veneer of objectivity	Trust calibration; uncertainty displays; performance metrics by group
	Confirmation bias	Belief-consistent information search	Disconfirmation prompts; counterfactual outputs; structured decision checklists
	Motivated social cognition	Identity protection; moral justification	Accountability requirements; independent human justification
	Availability & anchoring	Salience-driven judgment	Multiple or randomized outputs; delayed recommendations
	Individual differences in prejudice	Prejudiced attitudes; dominance motives	Restrict AI use in high-discretion contexts; targeted bias education
Feedback Loop	Biased human feedback	Reinforcement learning from biased behavior	Slowed retraining cycles; human review checkpoints; bias audits of feedback data

5. Breaking the Loop: Implications for bias reduction

Effective mitigation of algorithmic bias requires interventions that disrupt the psychological mechanisms that sustain bias across the human–AI loop, rather than isolated technical fixes. Below, we outline intervention strategies targeting human biases that influence both the creation and use of AI systems, with emphasis on insights from social and cognitive psychology.

A. Upstream interventions: Reducing Bias in Data Creation

Interventions to reduce bias in the creation of AI systems target the social-cognitive processes that govern data generation, data selection, and institutional priorities.

Promoting fair and diverse annotation. Because many AI systems rely on human-labeled data, reducing bias in annotation is critical. Recruiting demographically and culturally diverse annotators broadens the perspectives informing labels and reduces the dominance of majority viewpoints. Annotation protocols can also be improved through bias-awareness prompts, clear task instructions, and structured practice with feedback—conditions shown to reduce reliance on heuristics and stereotypes [67,68]. Practical factors such as fatigue, time pressure, distraction, and incentives should be actively managed, as these conditions increase reliance on cognitive shortcuts.

Although value-aligning measures will promote fairness, some degree of bias—particularly involving indirect expressions—may be unavoidable [74]. Transparency is therefore essential. Documenting annotator demographics, cultural context, and task instructions can clarify the

interpretive frame embedded in training labels and inform downstream model evaluation [130].

Reducing bias in dataset selection. Bias can also enter AI systems through the choice of training data. Dataset selection should therefore be guided by the intended use of the system. For example, AI tools designed for cross-cultural deployment require training data that adequately represent relevant populations. Importantly, proportional representation based on population frequency is often insufficient: achieving comparable performance across majority and minority groups typically requires equal representation of those groups in training data. This pattern is well documented in face classification models, where minority-class underrepresentation leads to persistent accuracy gaps [97]. In LLMs, careful curation of training data to remove human patterns of prejudice has also been shown to reduce bias in model performance [131].

Transparency is key here, too. Dataset characteristics—such as demographic composition, geographic origin, and historical context—should be clearly documented and communicated to developers and users [130]. Regulatory standards that mandate disclosure of training data properties would support these efforts.

Prioritizing fairness within institutional cultures. Organizational priorities play a central role in shaping algorithmic bias. In many technology firms, speed, scale, and profit have historically outweighed fairness considerations. Cultural change in corporations is challenging, but two forces appear particularly effective: sustained external pressure from researchers, advocacy groups, and users, and top-down regulation that mandates transparency, accountability, and bias mitigation [132]. Emerging legal frameworks, such as the EU's AI Act, illustrate how

structural incentives can normalize fairness as a core design requirement rather than an optional add-on. Progress will also depend on increasing gender, racial, and socioeconomic diversity within technology organizations, particularly in leadership roles where strategic priorities are set.

B. Downstream interventions: Reducing Bias in AI Use

Psychology is especially well positioned to inform interventions at the point of AI consumption, drawing on extensive research on prejudice reduction, judgment, and self-regulation. Evidence suggests that combining individual-level and structural interventions is most effective [133].

Individual-level interventions. Individual-level interventions involve educating users to detect and avoid the use of potentially biased AI outputs. Research on the self-regulation of prejudice identifies two core processes: detecting potential bias and implementing corrective responses [134,135]. Bias detection can be enhanced through education that emphasizes both the harms of bias and the difficulty of recognizing it, along with clear criteria for what constitutes biased output in the specific context of AI systems and clear instructions for how to avoid it. Once detected, unbiased decisions can be facilitated through the use of proactive strategies such as implementation intentions—if-then plans that link a situational cue to a specific action [136]—which can promote the prioritization of unbiased information [137].

Insights from judgment and decision-making research further suggest strategies to reduce overreliance on biased AI outputs. For example, encouraging users to consider alternative

hypotheses, generate counterfactuals, or explicitly justify decisions can reduce anchoring, confirmation bias, and undue trust in belief-consistent AI outputs [128,138].

Structural interventions. Structural interventions reduce bias by altering decision environments rather than relying on individual self-control. Classic examples include blind grading and blind auditions, which can successfully reduce bias by removing irrelevant social cues [139]. These approaches are effective precisely because they bypass the need for individuals to detect or regulate their own biases. Analogous strategies can be applied to AI use. For example, interface designs could reduce overreliance on single outputs by presenting multiple alternatives, displaying uncertainty estimates, restricting AI systems to clearly specified decision contexts and criteria, or requiring fact-checking against non-AI sources.

At the policy level, laws requiring transparency, documentation, and algorithmic impact assessments function as structural guardrails that shape how AI systems are interpreted and applied [140]. Such measures not only constrain technology but, by raising awareness, increase the public's awareness of how AI outputs are perceived, evaluated, and implemented.

6. A Psychology of Algorithmic Bias

More broadly, our analysis highlights the crucial role of psychology—its theories, methods, and perspectives—in understanding and addressing algorithmic bias. If biases expressed by AI systems originate in human cognition, motivation, and social structure, then algorithmic bias cannot be fully explained or mitigated without psychological science.

This perspective calls for deeper engagement of computer and data scientists with social and cognitive psychologists, particularly those studying prejudice, social cognition, and decision making. Psychology offers well-established frameworks for identifying bias, tracing its cognitive and motivational roots, and designing interventions that target both individual and structural processes in the context of AI technologies.

Of course, not all forms of algorithmic bias arise directly from human behavior: statistical artifacts, nonrepresentative sampling, model misspecification, and evaluation on restricted or unrepresentative test sets can introduce bias even in the absence of explicit human prejudice [74]. However, these technical errors are rarely independent from human judgment: as we have discussed, decisions about data selection, model tuning, evaluation criteria, and deployment contexts are themselves shaped by human goals, assumptions, and institutional incentives. A psychological approach therefore complements a technical analysis of algorithmic bias by clarifying how human decisions shape and interact with ostensibly technical failures.

Advancing a psychology of algorithmic bias will require structural changes in how AI research is conducted. It will require interdisciplinary teams that integrate psychological expertise with computational approaches from the earliest stages of question development and research design. Funding agencies can accelerate this shift by prioritizing research programs that explicitly link computational methods with theories of human cognition and social behavior. Similarly, training programs should prepare a new generation of researchers fluent in both psychological science and AI computation, capable of identifying bias across the full human–

machine loop. While such changes have already begun, particularly in the emergence of interdisciplinary data science centers, this approach can be expanded further to include social psychological expertise that specifically addresses the human sociocognitive processes that interact with AI.

7. Concluding remarks

In this article, we introduced the Human-AI Loop model of algorithmic bias, which conceptualizes bias in AI systems as emerging from dynamic, reciprocal interactions between humans and AI. By describing how human biases can enter during both the creation and use of AI systems, in a self-perpetuating cycle, this framework identifies human psychological processes as a root cause of algorithmic bias and explains its role in the persistence of algorithmic discrimination. Progress toward fairer AI will require a consideration of how these psychological processes shape the ways humans generate data, interpret algorithmic outputs, and incorporate AI into social decision-making, along with interventions that effectively target them.

Box 1. Algorithmic Implicit Bias & Bias Laundering (334/400 words)

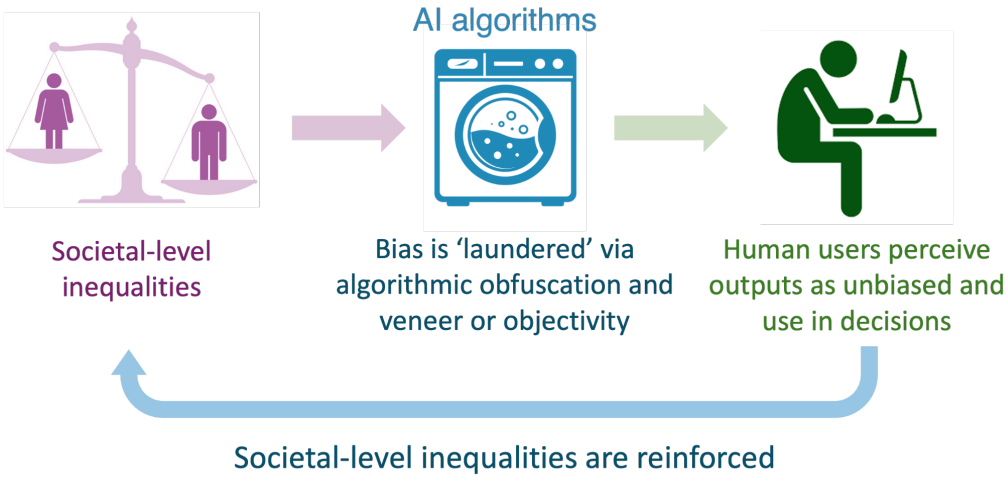
The "Human-AI Loop" model posits that algorithmic bias is not merely an artifact of computation, but a novel expression of human prejudice. That is, it reflects the transmission of prejudice between social systems and individuals via the interface of AI datasets and systems (Box 1 Fig 1). This perspective on AI reimagines the construct of implicit bias—the indirect expression of prejudice—as a multilevel process in which prejudice is communicated between individuals, social systems, and technology.

This view differs from past theories of implicit bias as an intra-individual process [141] or, more recently, as a reflection of structural inequalities on collective cognition [142,143]. In these cases, implicit bias exists within an individual, and its "implicitness" refers to how prejudice represented in one's mind is expressed indirectly in their behavior [144].

Algorithmic implicit bias, by contrast, represents the transmission of prejudice between completely independent entities—social structures (e.g., social disparities in training data), individuals (e.g., annotators and end-users), and AI computation—and therefore the indirect (i.e., implicit) nature of this transmission is distinctly discerned. Because the components of algorithmic bias are distinguished by level of analysis, each step in the production of implicit bias can be directly traced and quantified.

Traditional models suggest that implicit bias can operate within awareness or intention, but such claims have been difficult to establish empirically [144,145]. By contrast, the multi-

component nature of algorithmic implicit bias effectively precludes the roles of awareness or intention from the expression of prejudice. For this reason, this AI-mediated form of implicit bias is particularly dangerous: because its expression is obscured by the opacity of AI computation and "veneer of objectivity" with which AI outputs are consumed [93,117], its origins in human and societal prejudice are easily masked—a virtual 'laundering' of human bias.

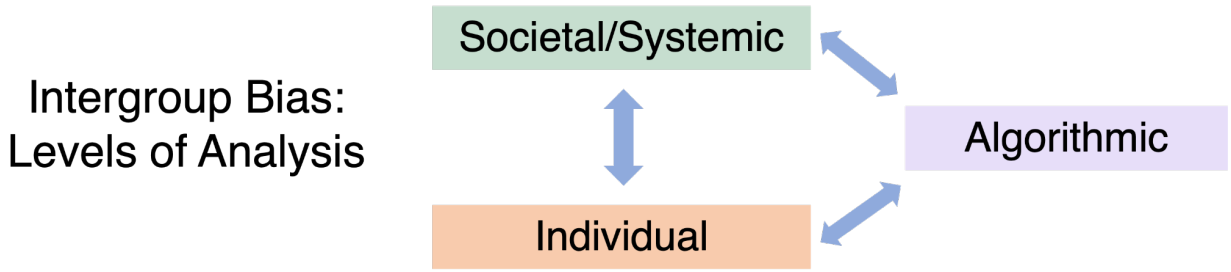


Box 1 Fig 1. The 'laundering' of human bias via AI systems

This reconceptualization of implicit bias—as a novel, computationally-mediated indirect expression—transforms our understanding of how prejudices are expressed and communicated without awareness or intention between humans and social systems. Future research is needed to characterize algorithmic implicit bias and understand its implications.

Box 2. The Computational Level of Analysis in Theories of Prejudice

Our analysis suggests an updated framework for theories of prejudice in the digital age. Traditionally, research on prejudice has focused on either the individual level of analysis, primarily within psychology, or the societal level, primarily within sociology, with limited crosstalk. The Human-AI Loop model introduces AI systems as a third, algorithmic (i.e., computational) level of analysis that functions as a unique mechanism through which biases are formed, represented, and expressed and which operates interactively with individual and systemic forms of prejudice.



Box 2 Fig 1. The algorithmic level of bias is distinct and complementary to individual and systemic levels.

AI systems learn, represent, and express social biases that resemble those observed in both individuals and social structures, yet they do so via distinct computational processes. Similar to individual-level forms of bias, AI systems are trained on collections of individual behaviors and judgments (e.g., text corpora, annotations), and they typically express bias through direct interactions with individual users that shape perceptions, beliefs, and decisions.

At the same time, algorithmic bias resembles societal-level prejudice in scale and breadth. Algorithmic models encode collective patterns of bias, often across large populations, akin to culturally-shared stereotypes that persist even when they are often rejected at the level of

individuals. Like other forms of systemic bias, AI systems enable the offloading of prejudice from individuals to opaque institutional structures, thereby allowing biased outcomes to be produced and justified with reduced personal accountability.

Due to its hybrid character, algorithmic bias constitutes a unique interface through which prejudice is transmitted between individual minds and social structures. By formalizing social bias in computational systems and reinserting it into everyday decision-making, AI systems have introduced a novel feedback pathway linking individual cognition, collective representations, and societal structures. By this perspective, understanding contemporary theories of prejudice will require treating technological systems not merely as reflections of human bias but as active contributors to a multilevel, recursive system of human prejudice.

Glossary

Algorithmic bias: systematic errors in AI systems that discriminate against disadvantaged social groups.

AI system: a computational system designed to perform tasks that typically require human intelligence, such as perception, language understanding, reasoning, or decision-making.

Social cognition: Mental processes that support how people perceive, interpret, judge, and act towards other people and social groups.

Prejudice: A positive or negative evaluation, attitude, or affective response toward a person based solely on their membership in a social group.

Stereotypes: Cognitive beliefs or expectations about the characteristics, traits, or behaviors of members of a social group.

Discrimination: differential treatment of individuals or groups based on their social category membership (e.g., race, gender, age, religion), rather than on their individual characteristics or merit.

Annotator bias: systematic distortions in human-labeled data that arise from annotators' beliefs, stereotypes, expectations, or situational influences, and which become embedded in the datasets used to train AI systems.

Perceptual hypodescent: Human tendency to perceive and classify multiracial faces as belonging to socially subordinate groups.

Gendered race effect. Biasing effect of racial stereotypes associated with male or female traits on human's perceptions and classifications of gender.

Other-race effect: The greater tendency for human perceivers to confuse or misidentify faces from other racial groups.

Proxy effect: When an AI system uses seemingly neutral input variables that are statistically correlated with protected attributes (e.g., race, gender, socioeconomic status), thereby indirectly reproducing group-based disparities even when the protected attribute itself is excluded.

Bias laundering: The process through which human sources of prejudice and discrimination become obscured in institutional or algorithm-based policies and decisions.

References

1. Gomez, C., Cho, S.M., Ke, S., Huang, C.-M., and Unberath, M. (2025). Human-AI collaboration is not very collaborative yet: a taxonomy of interaction patterns in AI-assisted decision making from a systematic review. Front. Comput. Sci. 6. https://doi.org/10.3389/fcomp.2024.1521066.

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

3. Noble, S.U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism (New York University Press) https://doi.org/10.18574/nyu/9781479833641.001.0001.

4. O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (Crown Publishing Group).

5. Zou, J., and Schiebinger, L. (2018). AI can be sexist and racist — it's time to make it fair. Nature 559, 324–326. https://doi.org/10.1038/d41586-018-05707-8.

6. Vlasceanu, M., and Amodio, D.M. (2022). Propagation of societal gender inequality by internet search algorithms. Proc. Natl. Acad. Sci. U.S.A. 119, e2204529119. https://doi.org/10.1073/pnas.2204529119.

7. Mattu, J.A., Jeff Larson,Lauren Kirchner,Surya (2016). What Algorithmic Injustice Looks Like in Real Life. ProPublica. https://www.propublica.org/article/what-algorithmic-injustice-looks-like-in-real-life.

8. Dastin, J. (2018). Insight - Amazon scraps secret AI recruiting tool that showed bias against women. Reuters.

9. Dodd, V. (2018). UK police use of facial recognition technology a failure, says report. The Guardian.

10. Hill, K. (2020). Wrongfully Accused by an Algorithm. The New York Times.

11. Tay, L., Woo, S.E., Hickman, L., Booth, B.M., and D'Mello, S. (2022). A Conceptual Framework for Investigating and Mitigating Machine-Learning Measurement Bias (MLMB) in Psychological Assessment. Advances in Methods and Practices in Psychological Science 5, 25152459211061337. https://doi.org/10.1177/25152459211061337.

12. Bolukbasi, T., Chang, K.-W., Zou, J.Y., Saligrama, V., and Kalai, A.T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In Advances in Neural Information Processing Systems (Curran Associates, Inc.).

13. Favaretto, M., De Clercq, E., and Elger, B.S. (2019). Big Data and discrimination: perils, promises and solutions. A systematic review. J Big Data 6, 12. https://doi.org/10.1186/s40537-019-0177-4.

14. Hardt, M., Price, E., and Srebro, N. (2016). Equality of Opportunity in Supervised Learning. In Advances in Neural Information Processing Systems (Curran Associates, Inc.).

15. Fiske, S.T. (1998). Stereotyping, prejudice, and discrimination. In The handbook of social psychology, Vols. 1-2, 4th ed (McGraw-Hill), pp. 357–411.

16. Kawakami, K., Amodio, D.M., and Hugenberg, K. (2017). Intergroup perception and cognition: An integrative framework for understanding the causes and consequences of social categorization. In Advances in experimental social psychology Advances in experimental social psychology. (Elsevier Academic Press), pp. 1–80. https://doi.org/10.1016/bs.aesp.2016.10.001.

17. Williams, N.W., Casas, A., Aslett, K., and Wilkerson, J. (2025). When Conservatives See Red but Liberals Feel Blue: Labeler Characteristics and Variation in Content Annotation. The Journal of Politics, 000–000. https://doi.org/10.1086/735397.

18. Chen, J.M. (2019). An integrative review of impression formation processes for multiracial individuals. Social & Personality Psych 13, e12430. https://doi.org/10.1111/spc3.12430.

19. Ho, A.K., Kteily, N.S., and Chen, J.M. (2017). “You’re one of us”: Black Americans’ use of hypodescent and its association with egalitarianism. J Pers Soc Psychol 113, 753–768. https://doi.org/10.1037/pspi0000107.

20. Peery, D., and Bodenhausen, G.V. (2008). Black + white = black: hypodescent in reflexive categorization of racially ambiguous faces. Psychol Sci 19, 973–977. https://doi.org/10.1111/j.1467-9280.2008.02185.x.

21. Krosch, A.R., Berntsen, L., Amodio, D.M., Jost, J.T., and Van Bavel, J.J. (2013). On the ideology of hypodescent: Political conservatism predicts categorization of racially ambiguous faces as Black. Journal of Experimental Social Psychology 49, 1196–1203. https://doi.org/10.1016/j.jesp.2013.05.009.

22. Krosch, A.R., and Amodio, D.M. (2014). Economic scarcity alters the perception of race. Proc. Natl. Acad. Sci. U.S.A. 111, 9079–9084. https://doi.org/10.1073/pnas.1404448111.

23. Chen, J.M., and Hamilton, D.L. (2012). Natural ambiguities: Racial categorization of multiracial individuals. Journal of Experimental Social Psychology 48, 152–164. https://doi.org/10.1016/j.jesp.2011.10.005.

24. Galinsky, A.D., Hall, E.V., and Cuddy, A.J.C. (2013). Gendered Races: Implications for Interracial Marriage, Leadership Selection, and Athletic Participation. Psychol Sci 24, 498–506. https://doi.org/10.1177/0956797612457783.

25. Johnson, K.L., Freeman, J.B., and Pauker, K. (2012). Race is gendered: How covarying phenotypes and stereotypes bias sex categorization. Journal of Personality and Social Psychology 102, 116–131. https://doi.org/10.1037/a0025335.

26. Dotsch, R., Wigboldus, D.H.J., Langner, O., and van Knippenberg, A. (2008). Ethnic Out-Group Faces Are Biased in the Prejudiced Mind. Psychol Sci 19, 978–980. https://doi.org/10.1111/j.1467-9280.2008.02186.x.

27. Halberstadt, A.G., Cooke, A.N., Garner, P.W., Hughes, S.A., Oertwig, D., and Neupert, S.D. (2022). Racialized emotion recognition accuracy and anger bias of children's faces. Emotion 22, 403–417. https://doi.org/10.1037/emo0000756.

28. Plant, E.A., Hyde, J.S., Keltner, D., and Devine, P.G. (2000). The Gender Stereotyping of Emotions. Psychology of Women Quarterly 24, 81–92. https://doi.org/10.1111/j.1471-6402.2000.tb01024.x.

29. Hess, U., Adams, R.B., Jr., Grammer, K., and Kleck, R.E. (2009). Face gender and emotion expression: Are angry women more like men? Journal of Vision 9, 19. https://doi.org/10.1167/9.12.19.

30. Adams, R.B., Albohn, D.N., Hedgecoth, N., Garrido, C.O., and Adams, K.D. (2022). Angry White Faces: A Contradiction of Racial Stereotypes and Emotion-Resembling Appearance. Affec Sci 3, 46–61. https://doi.org/10.1007/s42761-021-00091-5.

31. Brooks, J.A., Stolier, R.M., and Freeman, J.B. (2018). Stereotypes Bias Visual Prototypes for Sex and Emotion Categories. Social Cognition 36, 481–493. https://doi.org/10.1521/soco.2018.36.5.481.

32. Chen, Y., and Joo, J. (2021). Understanding and Mitigating Annotation Bias in Facial Expression Recognition. In, pp. 14980–14991.

33. Meissner, C.A., and Brigham, J.C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. Psychology, Public Policy, and Law 7, 3–35. https://doi.org/10.1037/1076-8971.7.1.3.

34. Hugenberg, K., Young, S.G., Bernstein, M.J., and Sacco, D.F. (2010). The categorization-individuation model: An integrative account of the other-race recognition deficit. Psychological Review 117, 1168–1187. https://doi.org/10.1037/a0020463.

35. Ho, A.K., Roberts, S.O., and Gelman, S.A. (2015). Essentialism and Racial Bias Jointly Contribute to the Categorization of Multiracial Individuals. Psychol Sci 26, 1639–1645. https://doi.org/10.1177/0956797615596436.

36. Casey, J.P., Vanman, E.J., and Barlow, F.K. (2025). Empathic Conservatives and Moralizing Liberals: Political Intergroup Empathy Varies by Political Ideology and Is Explained by Moral Judgment. Pers Soc Psychol Bull 51, 678–700. https://doi.org/10.1177/01461672231198001.

37. Roussos, G., and Dovidio, J.F. (2018). Hate Speech Is in the Eye of the Beholder: The Influence of Racial Attitudes and Freedom of Speech Beliefs on Perceptions of Racially

Motivated Threats of Violence. Social Psychological and Personality Science 9, 176–185. https://doi.org/10.1177/1948550617748728.

38. White II, M.H., and Crandall, C.S. (2017). Freedom of racist speech: Ego and expressive threats. Journal of Personality and Social Psychology 113, 413–429. https://doi.org/10.1037/pspi0000095.

39. Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N.A. (2019). The Risk of Racial Bias in Hate Speech Detection. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, A. Korhonen, D. Traum, and L. Màrquez, eds. (Association for Computational Linguistics), pp. 1668–1678. https://doi.org/10.18653/v1/P19-1163.

40. Davani, A.M., Atari, M., Kennedy, B., and Dehghani, M. (2023). Hate Speech Classifiers Learn Normative Social Stereotypes. Transactions of the Association for Computational Linguistics 11, 300–319. https://doi.org/10.1162/tacl_a_00550.

41. Das, A., Zhang, Z., Hasan, N., Sarkar, S., Jamshidi, F., Bhattacharya, T., Rahgouy, M., Raychawdhary, N., Feng, D., Jain, V., et al. (2024). Investigating Annotator Bias in Large Language Models for Hate Speech Detection. Preprint at arXiv, https://doi.org/10.48550/arXiv.2406.11109 https://doi.org/10.48550/arXiv.2406.11109.

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

43. Chen, Z. (2023). Ethics and discrimination in artificial intelligence-enabled recruitment practices. Humanit Soc Sci Commun 10, 1–12. https://doi.org/10.1057/s41599-023-02079-x.

44. Wilson, K., and Caliskan, A. (2024). Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 7, 1578–1590. https://doi.org/10.1609/aies.v7i1.31748.

45. Heilman, M.E. (2012). Gender stereotypes and workplace bias. Research in Organizational Behavior 32, 113–135. https://doi.org/10.1016/j.riob.2012.11.003.

46. Quillian, L., Pager, D., Hexel, O., and Midtbøen, A.H. (2017). Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. Proceedings of the National Academy of Sciences 114, 10870–10875. https://doi.org/10.1073/pnas.1706255114.

47. Steinpreis, R.E., Anders, K.A., and Ritzke, D. (1999). The Impact of Gender on the Review of the Curricula Vitae of Job Applicants and Tenure Candidates: A National Empirical Study. Sex Roles 41, 509–528. https://doi.org/10.1023/A:1018839203698.

48. Bertrand, M., and Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. American Economic Review 94, 991–1013. https://doi.org/10.1257/0002828042002561.

49. Zavala, V.A., Bracci, P.M., Carethers, J.M., Carvajal-Carmona, L., Coggins, N.B., Cruz-Correa, M.R., Davis, M., de Smith, A.J., Dutil, J., Figueiredo, J.C., et al. (2021). Cancer health disparities in racial/ethnic minorities in the United States. Br J Cancer 124, 315–332. https://doi.org/10.1038/s41416-020-01038-6.

50. Penner, L.A., Dovidio, J.F., Gonzalez, R., Albrecht, T.L., Chapman, R., Foster, T., Harper, F.W.K., Hagiwara, N., Hamel, L.M., Shields, A.F., et al. (2016). The Effects of Oncologist Implicit Racial Bias in Racially Discordant Oncology Interactions. JCO 34, 2874–2880. https://doi.org/10.1200/JCO.2015.66.3658.

51. Green, A.R., Carney, D.R., Pallin, D.J., Ngo, L.H., Raymond, K.L., Iezzoni, L.I., and Banaji, M.R. (2007). Implicit Bias among Physicians and its Prediction of Thrombolysis Decisions for Black and White Patients. J GEN INTERN MED 22, 1231–1238. https://doi.org/10.1007/s11606-007-0258-5.

52. Hoffman, K.M., Trawalter, S., Axt, J.R., and Oliver, M.N. (2016). Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. Proceedings of the National Academy of Sciences 113, 4296–4301. https://doi.org/10.1073/pnas.1516047113.

53. Johnson, R.L., Roter, D., Powe, N.R., and Cooper, L.A. (2004). Patient Race/Ethnicity and Quality of Patient–Physician Communication During Medical Visits. Am J Public Health 94, 2084–2090. https://doi.org/10.2105/AJPH.94.12.2084.

54. Celi, L.A., Cellini, J., Charpignon, M.-L., Dee, E.C., Dernoncourt, F., Eber, R., Mitchell, W.G., Moukheiber, L., Schirmer, J., Situ, J., et al. (2022). Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. PLOS Digital Health 1, e0000022. https://doi.org/10.1371/journal.pdig.0000022.

55. Markowitz, D.M. (2022). Gender and ethnicity bias in medicine: a text analysis of 1.8 million critical care records. PNAS Nexus 1. https://doi.org/10.1093/pnasnexus/pgac157.

56. Chen, R.J., Wang, J.J., Williamson, D.F.K., Chen, T.Y., Lipkova, J., Lu, M.Y., Sahai, S., and Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat. Biomed. Eng 7, 719–742. https://doi.org/10.1038/s41551-023-01056-8.

57. Timmons, A.C., Duong, J.B., Simo Fiallo, N., Lee, T., Vo, H.P.Q., Ahle, M.W., Comer, J.S., Brewer, L.C., Frazier, S.L., and Chaspari, T. (2023). A Call to Action on Assessing and Mitigating Bias in Artificial Intelligence Applications for Mental Health. Perspect Psychol Sci 18, 1062–1096. https://doi.org/10.1177/17456916221134490.

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

59. Ding, Y., You, J., Machulla, T.-K., Jacobs, J., Sen, P., and Höllerer, T. (2022). Impact of Annotator Demographics on Sentiment Dataset Labeling. Proc. ACM Hum.-Comput. Interact. 6, 519:1-519:22. https://doi.org/10.1145/3555632.

60. Carvacho, H., Zick, A., Haye, A., González, R., Manzi, J., Kocik, C., and Bertl, M. (2013). On the relation between social class and prejudice: The roles of education, income, and ideological attitudes. European Journal of Social Psychology 43, 272–285. https://doi.org/10.1002/ejsp.1961.

61. Franssen, V., Dhont, K., and Hiel, A.V. (2013). Age-Related Differences in Ethnic Prejudice: Evidence of the Mediating Effect of Right-Wing Attitudes. Journal of Community & Applied Social Psychology 23, 252–257. https://doi.org/10.1002/casp.2109.

62. Navarrete, C.D., McDonald, M.M., Molina, L.E., and Sidanius, J. (2010). Prejudice at the nexus of race and gender: An outgroup male target hypothesis. Journal of Personality and Social Psychology 98, 933–945. https://doi.org/10.1037/a0017931.

63. Kunda, Z., and Spencer, S.J. (2003). When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and application. Psychological Bulletin 129, 522–544. https://doi.org/10.1037/0033-2909.129.4.522.

64. Paolini, S., Crisp, R.J., and McIntyre, K. (2009). Accountability moderates member-to-group generalization: Testing a dual process model of stereotype change. Journal of Experimental Social Psychology 45, 676–685. https://doi.org/10.1016/j.jesp.2009.03.005.

65. McDANIEL, M.A., Hartman, N.S., Whetzel, D.L., and Grubb Iii, W.L. (2007). Situational Judgment Tests, Response Instructions, and Validity: A Meta-Analysis. Personnel Psychology 60, 63–91. https://doi.org/10.1111/j.1744-6570.2007.00065.x.

66. Parmar, M., Mishra, S., Geva, M., and Baral, C. (2024). Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions. Preprint at arXiv, https://doi.org/10.48550/arXiv.2205.00415 https://doi.org/10.48550/arXiv.2205.00415.

67. Devine, P.G., and Elliot, A.J. (1995). Are Racial Stereotypes Really Fading? The Princeton Trilogy Revisited. Pers Soc Psychol Bull 21, 1139–1150. https://doi.org/10.1177/01461672952111002.

68. Hsieh, G., and Kocielnik, R. (2016). You Get Who You Pay for: The Impact of Incentives on Participation Bias. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing CSCW '16. (Association for Computing Machinery), pp. 823–835. https://doi.org/10.1145/2818048.2819936.

69. Kashima, Y., Fiedler, K., and Freytag, P. (2008). Stereotype Dynamics: Language-based Approaches to the Formation, Maintenance, and Transformation of Stereotypes (Taylor & Francis).

70. Maass, A., Salvi, D., Arcuri, L., and Semin, G.R. (1989). Language use in intergroup contexts: The linguistic intergroup bias. Journal of Personality and Social Psychology 57, 981–993. https://doi.org/10.1037/0022-3514.57.6.981.

71. Rhodes, M., Gelman, S.A., and Leslie, S.-J. (2025). How generic language shapes the development of social thought. Trends in Cognitive Sciences 29, 122–132. https://doi.org/10.1016/j.tics.2024.09.012.

72. Bai, X., Wang, A., Sucholutsky, I., and Griffiths, T.L. (2025). Explicitly unbiased large language models still form biased associations. Proceedings of the National Academy of Sciences 122, e2416228122. https://doi.org/10.1073/pnas.2416228122.

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

74. Waldfogel, J. (2025). The Welfare Effects of Gender-Inclusive Intellectual Property Creation: Evidence from Books. Journal of Political Economy 133, 2229–2264. https://doi.org/10.1086/734876.

75. Hovy, D., and Søgaard, A. (2015). Tagging Performance Correlates with Author Age. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), C. Zong and M. Strube, eds. (Association for Computational Linguistics), pp. 483–488. https://doi.org/10.3115/v1/P15-2079.

76. Wang, A., Morgenstern, J., and Dickerson, J.P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. Nat Mach Intell 7, 400–411. https://doi.org/10.1038/s42256-025-00986-z.

77. Henrich, J., Heine, S.J., and Norenzayan, A. (2010). The weirdest people in the world? Behavioral and Brain Sciences 33, 61–83. https://doi.org/10.1017/S0140525X0999152X.

78. Date, S., Deshmukh, S.N., Boyd, R., Ashokkumar, A., and Pennebaker, J.W. (2024). Designing of a Novel Framework for Marathi Natural Language Processing: MR-LIWC2015. International Journal of Intelligent Systems and Applications in Engineering 12, 01–14.

79. Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning Word Vectors for 157 Languages. Preprint at arXiv, https://doi.org/10.48550/arXiv.1802.06893 https://doi.org/10.48550/arXiv.1802.06893.

80. Toney-Wails, A., and Caliskan, A. (2021). ValNorm Quantifies Semantics to Reveal Consistent Valence Biases Across Languages and Over Centuries. Preprint at arXiv, https://doi.org/10.48550/arXiv.2006.03950 https://doi.org/10.48550/arXiv.2006.03950.

81. Jackson, J.C., Liu, Y., Wang, Z., and Brady, W.J. (2026). Large AI Models Have a Prioritization Problem: Policy Implications and Solutions. Policy Insights from the Behavioral and Brain Sciences 13, 5–13. https://doi.org/10.1177/23727322251408311.

82. Sourati, Z., Karimi-Malekabadi, F., Ozcan, M., McDaniel, C., Ziabari, A., Trager, J., Tak, A., Chen, M., Morstatter, F., and Dehghani, M. (2025). The Shrinking Landscape of Linguistic Diversity in the Age of Large Language Models. Preprint at arXiv, https://doi.org/10.48550/arXiv.2502.11266 https://doi.org/10.48550/arXiv.2502.11266.

83. Charlesworth, T.E.S., Sanjeev, N., Hatzenbuehler, M.L., and Banaji, M.R. (2023). Identifying and predicting stereotype change in large language corpora: 72 groups, 115 years (1900–2015), and four text sources. Journal of Personality and Social Psychology 125, 969–990. https://doi.org/10.1037/pspa0000354.

84. Bailey, A.H., Williams, A., and Cimpian, A. (2022). Based on billions of words on the internet, people = men. Science Advances 8, eabm2463. https://doi.org/10.1126/sciadv.abm2463.

85. Cheryan, S., and Markus, H.R. (2020). Masculine defaults: Identifying and mitigating hidden cultural biases. Psychological Review 127, 1022–1052. https://doi.org/10.1037/rev0000209.

86. Brady, W.J., Jackson, J.C., Lindström, B., and Crockett, M.J. (2023). Algorithm-mediated social learning in online social networks. Trends in Cognitive Sciences 27, 947–960. https://doi.org/10.1016/j.tics.2023.06.008.

87. Buolamwini, J., and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (PMLR), pp. 77–91.

88. Karkkainen, K., and Joo, J. (2021). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In, pp. 1548–1558.

89. Morales, A., Fierrez, J., Vera-Rodriguez, R., and Tolosana, R. (2021). SensitiveNets: Learning Agnostic Representations with Application to Face Images. IEEE Transactions on Pattern Analysis and Machine Intelligence 43, 2158–2164. https://doi.org/10.1109/TPAMI.2020.3015420.

90. Lewis, M., and Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. Nat Hum Behav 4, 1021–1028. https://doi.org/10.1038/s41562-020-0918-6.

91. Charlesworth, T.E.S., Yang, V., Mann, T.C., Kurdi, B., and Banaji, M.R. (2021). Gender Stereotypes in Natural Language: Word Embeddings Show Robust Consistency Across Child and Adult Language Corpora of More Than 65 Million Words. Psychol Sci 32, 218–240. https://doi.org/10.1177/0956797620963619.

92. Card, D., Chang, S., Becker, C., Mendelsohn, J., Voigt, R., Boustan, L., Abramitzky, R., and Jurafsky, D. (2022). Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration. Proceedings of the National Academy of Sciences 119, e2120510119. https://doi.org/10.1073/pnas.2120510119.

93. Kozlowski, A.C., Taddy, M., and Evans, J.A. (2019). The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. Am Sociol Rev 84, 905–949. https://doi.org/10.1177/0003122419877135.

94. Charlesworth, T.E.S., and Hatzenbuehler, M.L. (2025). The Stigma Stability Framework: An Integrated Theory of How and Why Society Transmits Stigma Across History. Social and Personality Psychology Compass 19, e70051. https://doi.org/10.1111/spc3.70051.

95. Kordzadeh, N., and Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. European Journal of Information Systems 31, 388–409. https://doi.org/10.1080/0960085X.2021.1927212.

96. Neil, R., and Zanger-Tishler, M. (2025). Algorithmic Bias in Criminal Risk Assessment: The Consequences of Racial Differences in Arrest as a Measure of Crime. Annual Review of Criminology 8, 97–119. https://doi.org/10.1146/annurev-criminol-022422-125019.

97. Fogliato, R., Xiang, A., Lipton, Z., Nagin, D., and Chouldechova, A. (2021). On the Validity of Arrest as a Proxy for Offense: Race and the Likelihood of Arrest for Violent Crimes. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society AIES '21. (Association for Computing Machinery), pp. 100–111. https://doi.org/10.1145/3461702.3462538.

98. Brady, W.J., Crockett, M.J., and Van Bavel, J.J. (2020). The MAD Model of Moral Contagion: The Role of Motivation, Attention, and Design in the Spread of Moralized Content Online. Perspect Psychol Sci 15, 978–1010. https://doi.org/10.1177/1745691620917336.

99. Brady, W.J., McLoughlin, K.L., Torres, M.P., Luo, K.F., Gendron, M., and Crockett, M.J. (2023). Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility. Nat Hum Behav 7, 917–927. https://doi.org/10.1038/s41562-023-01582-0.

100. Seyyed-Kalantari, L., Zhang, H., McDermott, M.B.A., Chen, I.Y., and Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 27, 2176–2182. https://doi.org/10.1038/s41591-021-01595-0.

101. Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., and Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proceedings of the National Academy of Sciences 117, 12592–12594. https://doi.org/10.1073/pnas.1919012117.

102. McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., et al. (2020). International evaluation of an AI system for breast cancer screening. Nature 577, 89–94. https://doi.org/10.1038/s41586-019-1799-6.

103. Norori, N., Hu, Q., Aellen, F.M., Faraci, F.D., and Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. Patterns 2. https://doi.org/10.1016/j.patter.2021.100347.

104. Charlesworth, T.E.S., Caliskan, A., and Banaji, M.R. (2022). Historical representations of social groups across 200 years of word embeddings from Google Books. Proceedings of the National Academy of Sciences 119, e2121798119. https://doi.org/10.1073/pnas.2121798119.

105. U.S. Equal Employment Opportunity Commission (2024). High Tech, Low Inclusion Diversity in the High Tech Workforce and Sector 2014-2022.

106. Ely, R.J., and Thomas, D.A. (2001). Cultural Diversity at Work: The Effects of Diversity Perspectives on Work Group Processes and Outcomes. Administrative Science Quarterly 46, 229–273. https://doi.org/10.2307/2667087.

107. Hebl, M., Cheng, S.K., and Ng, L.C. (2020). Modern Discrimination in Organizations. Annual Review of Organizational Psychology and Organizational Behavior 7, 257–282. https://doi.org/10.1146/annurev-orgpsych-012119-044948.

108. The Reason This “Racist Soap Dispenser” Doesn’t Work on Black Skin (2015). Mic. https://www.mic.com/articles/124899/the-reason-this-racist-soap-dispenser-doesn-t-work-on-black-skin.

109. Gray, M.L., and Suri, S. Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass - Mary L. Gray, Siddharth Suri - Google Books. https://books.google.nl/books?hl=en&lr=&id=8AmXDwAAQBAJ&oi=fnd&pg=PP1&ots=WVN_PX7X0q&sig=QdVm8E4-gn7T9sR--gCEMUUIHPk&redir_esc=y#v=onepage&q&f=false.

110. Raji, I.D., Costanza-Chock, S., and Buolamwini, J. Change from the outside: towards credible third-party audits of AI systems. In Missing Links in AI Governance (United Nations Educational, Scientific and Cultural Organization (UNESCO)), pp. 5–26.

111. Drissi, M. (2024). More is Less? A Simulation-Based Approach to Dynamic Interactions between Biases in Multimodal Models. Preprint at arXiv, https://doi.org/10.48550/arXiv.2412.17505 https://doi.org/10.48550/arXiv.2412.17505.

112. Dikmen, M., and Burns, C. (2022). The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. International Journal of Human-Computer Studies 162, 102792. https://doi.org/10.1016/j.ijhcs.2022.102792.

113. Schultheiß, S., and Lewandowski, D. (2023). Misplaced trust? The relationship between trust, ability to identify commercially influenced results and search engine preference. Journal of Information Science 49, 609–623. https://doi.org/10.1177/01655515211014157.

114. Girouard-Hallam, L.N., and Danovitch, J.H. (2025). Children's trust in Google's ability to answer questions about the past, present, and future. Computers in Human Behavior 165, 108496. https://doi.org/10.1016/j.chb.2024.108496.

115. Stephany, F., and Duszynski, J. (2026). Women Worry, Men Adopt: How Gendered Perceptions Shape the Use of Generative AI. Preprint at arXiv, https://doi.org/10.48550/arXiv.2601.03880 https://doi.org/10.48550/arXiv.2601.03880.

116. Choung, H., David ,Prabu, and and Ross, A. (2023). Trust in AI and Its Role in the Acceptance of AI Technologies. International Journal of Human–Computer Interaction 39, 1727–1739. https://doi.org/10.1080/10447318.2022.2050543.

117. Schoeffer, J., Kuehl, N., and Machowski, Y. (2022). "There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency FAccT '22. (Association for Computing Machinery), pp. 1616–1628. https://doi.org/10.1145/3531146.3533218.

118. Bigman, Y.E., and Gray, K. (2018). People are averse to machines making moral decisions. Cognition 181, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003.

119. Myers, S., and Everett, J.A.C. (2025). People expect artificial moral advisors to be more utilitarian and distrust utilitarian moral advisors. Cognition 256, 106028. https://doi.org/10.1016/j.cognition.2024.106028.

120. Verma, A., Lamsal, K., and Verma, P. (2022). An investigation of skill requirements in artificial intelligence and machine learning job advertisements. Industry and Higher Education 36, 63–73. https://doi.org/10.1177/0950422221990990.

121. Kunda, Z. (1990). The case for motivated reasoning. Psychological Bulletin 108, 480–498. https://doi.org/10.1037/0033-2909.108.3.480.

122. Sidanius, J., and Pratto, F. (1999). Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression (Cambridge University Press).

123. Araujo, T., Helberger, N., Kruikemeier, S., and de Vreese, C.H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI & Soc 35, 611–623. https://doi.org/10.1007/s00146-019-00931-w.

124. Bertrand, A., Belloum, R., Eagan, J.R., and Maxwell, W. (2022). How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. In Proceedings of the 2022 AAAI/ACM

Conference on AI, Ethics, and Society AIES '22. (Association for Computing Machinery), pp. 78–91. https://doi.org/10.1145/3514094.3534164.

125. Glikson, E., and Woolley, A.W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. ANNALS 14, 627–660. https://doi.org/10.5465/annals.2018.0057.

126. Langer, M., and Landers, R.N. (2021). The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. Computers in Human Behavior 123, 106878. https://doi.org/10.1016/j.chb.2021.106878.

127. Tversky, A., and Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. Science 185, 1124–1131. https://doi.org/10.1126/science.185.4157.1124.

128. Rastogi, C., Zhang, Y., Wei, D., Varshney, K.R., Dhurandhar, A., and Tomsett, R. (2022). Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. Proc. ACM Hum.-Comput. Interact. 6, 83:1-83:22. https://doi.org/10.1145/3512930.

129. Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. Ethics Inf Technol 20, 5–14. https://doi.org/10.1007/s10676-017-9430-8.

130. Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S.R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S.R., et al. (2025). Towards Understanding Sycophancy in Language Models. Preprint at arXiv, https://doi.org/10.48550/arXiv.2310.13548 https://doi.org/10.48550/arXiv.2310.13548.

131. Hu, T., Kyrychenko, Y., Rathje, S., Collier, N., van der Linden, S., and Roozenbeek, J. (2025). Generative language models exhibit social identity biases. Nat Comput Sci 5, 65–75. https://doi.org/10.1038/s43588-024-00741-1.

132. Devine, P.G. (1989). Stereotypes and prejudice: Their automatic and controlled components. Journal of Personality and Social Psychology 56, 5–18. https://doi.org/10.1037/0022-3514.56.1.5.

133. Amodio, D.M., Harmon-Jones, E., Devine, P.G., Curtin, J.J., Hartley, S.L., and Covert, A.E. (2004). Neural Signals for the Detection of Unintentional Race Bias. Psychol Sci 15, 88–93. https://doi.org/10.1111/j.0963-7214.2004.01502003.x.

134. Sheeran, P., Listrom, O., and Gollwitzer, P.M. (2025). The when and how of planning: Meta-analysis of the scope and components of implementation intentions in 642 tests. European Review of Social Psychology 36, 162–194. https://doi.org/10.1080/10463283.2024.2334563.

135. Ha, T., and Kim, S. (2024). Improving Trust in AI with Mitigating Confirmation Bias: Effects of Explanation Type and Debiasing Strategy for Decision-Making with Explainable AI.

International Journal of Human–Computer Interaction 40, 8562–8573. https://doi.org/10.1080/10447318.2023.2285640.

136. Goldin, C., and Rouse, C. (2000). Orchestrating Impartiality: The Impact of “Blind” Auditions on Female Musicians. American Economic Review 90, 715–741. https://doi.org/10.1257/aer.90.4.715.

137. Greenwald, A.G., and Banaji, M.R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. Psychological Review 102, 4–27. https://doi.org/10.1037/0033-295X.102.1.4.

138. Payne, B.K., Vuletich, H.A., and Lundberg, K.B. (2017). The Bias of Crowds: How Implicit Bias Bridges Personal and Systemic Prejudice. Psychological Inquiry 28, 233–248. https://doi.org/10.1080/1047840X.2017.1335568.

139. Schultner, D.T., Stillerman, B.S., Lindström, B.R., Hackel, L.M., Hagen, D.R., Jostmann, N.B., and Amodio, D.M. (2024). Transmission of societal stereotypes to individual-level prejudice through instrumental learning. Proceedings of the National Academy of Sciences 121, e2414518121. https://doi.org/10.1073/pnas.2414518121.

140. Amodio, D.M. (2025). A learning and memory account of impression formation and updating. Nat Rev Psychol 4, 417–432. https://doi.org/10.1038/s44159-025-00445-x.

141. Corneille, O., and Hütter, M. (2020). Implicit? What Do You Mean? A Comprehensive Review of the Delusive Implicitness Construct in Attitude Research. Pers Soc Psychol Rev 24, 212–232. https://doi.org/10.1177/1088868320911325.