****PREPRINT****

August 7, 2026

The Psychology of Algorithmic Bias

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

[1]Department of Psychology, University of Amsterdam

[2]Kellogg School of Management, Northwestern University

Correspondence: david.amodio@gmail.com

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

Abstract

Algorithmic bias—the systematic discrimination of disadvantaged social groups by AI systems—is widely documented, yet its psychological foundations remain poorly understood. We argue that algorithmic bias originates primarily from human prejudice and social cognition, which shape AI training data and use of AI systems. We introduce the psychology-centered *Human–AI Loop* model that explains how biases can enter an AI system during its creation and consumption, specifying the psychological mechanisms operating at each entry point. These mechanisms and their interactions create feedback loops through which AI systems amplify, perpetuate, and obscure existing prejudices and social inequalities. We then outline psychology-informed strategies for disrupting these cycles and discuss implications for theories of prejudice and discrimination in the age of AI.

Bias in the Machine

From the use of large language models (LLMs) to facial recognition systems and 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 the 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 people with darker skin [9,10]. These error patterns in AI systems are not random [11]; they reflect systematic 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 remain limited by an incomplete account of its fundamental source—humans—and the psychological and behavioral pathways through which human biases infiltrate AI.

In this article, we argue that algorithmic bias is rooted in human **social cognition** and **prejudice**—processes that shape nearly every stage of the AI lifecycle. AI systems are conceived, trained, deployed, interpreted, and used by humans whose judgments reflect well-

established biases in perception, evaluation, and decision making. Explaining and mitigating algorithmic bias therefore requires understanding the psychological mechanisms through which human biases interact with AI systems.

A Psychological Perspective on Algorithmic Bias

Algorithmic bias emerges broadly from AI computation and its interaction with humans during an AI system's creation and use. Whereas prior work has focused on computational sources of algorithmic bias [2], our focus here is on the human contribution: we argue that a comprehensive account of algorithmic bias requires a psychological perspective grounded in social cognition and behavioral science in addition to computational perspectives.

From a psychological perspective, algorithmic bias is not merely a computational artifact, but can be viewed as a novel expression of human prejudice via algorithmic computation. That is, it represents the transmission of prejudice between individuals and AI systems via training data and model tuning, which in turn generate outputs that reinforce, legitimize, and propagate existing social inequalities in users and society—a self-perpetuating loop of algorithmic bias [5,15].

Previous "loop" models have described algorithmic bias as emerging from interactions among users, algorithms, and society [15–18], consistent with broader sociotechnical perspectives [19]. However, they have not specified the psychological mechanisms through which human

biases enter and are transmitted by AI systems—processes essential to the formation and expression of algorithmic bias.

To this end, we propose a Human–AI Loop model (Figure 1, Key Figure), which emphasizes human contributions to algorithmic bias. This model identifies two primary entry points through which human prejudice influences AI: during AI creation, via the production and curation of training data, and during AI consumption, via users' interpretation and application of AI outputs. By specifying the psychological mechanisms operating at each stage, the model offers a more complete account of how algorithmic bias is formed, perpetuated, and ultimately mitigated.

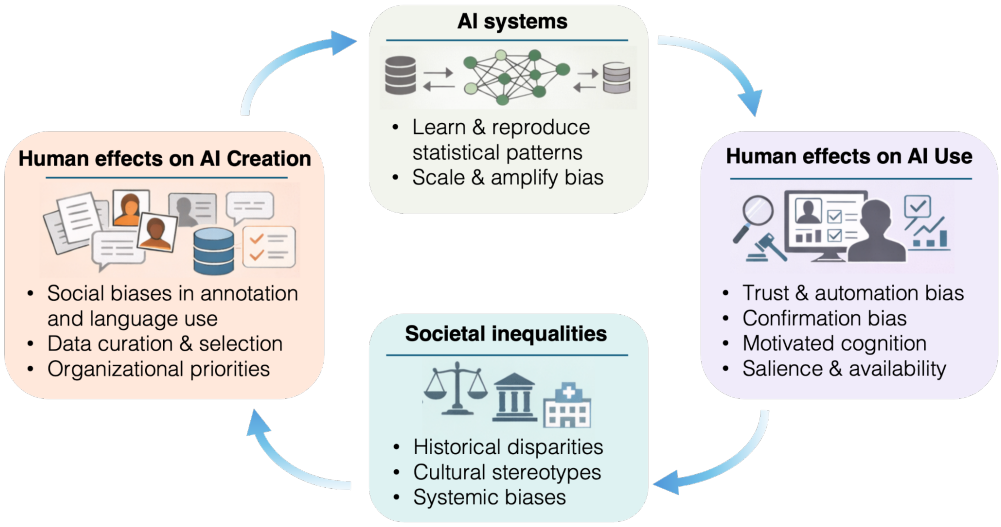


Figure 1. The "Human-AI Loop" model of algorithmic bias. Human psychological processes contribute to algorithmic bias during AI creation and during AI consumption. These human influences interact with computational processes to produce biased algorithmic outputs, which shape human judgments and behaviors, generate new training data, and ultimately reinforce—and potentially amplify—existing societal inequalities through a self-reinforcing Human–AI Loop.

In what follows, we describe major components of the Human-AI Loop and the psychological mechanisms through which prejudice infiltrates AI systems at each point of human-AI interaction. We then outline approaches for breaking this cycle, with an emphasis on psychology-based interventions, and discuss how our analysis advances psychological theories of prejudice in the digital age.

Human Sources of Bias in AI Creation

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.

Stereotype bias effects on categorization, judgment, and decision making

Many AI systems—including facial recognition, content moderation, hiring, and medical diagnostic tools—are trained on data labeled by human annotators. Because annotation often requires subjective judgments, it is susceptible to the same stereotypes and prejudices that shape human perception and decision making [20,21], producing **annotator bias**. As a result, these biases become embedded in training data and reproduced by AI systems [22]. Below, we describe a set of well-established social-cognitive effects likely to influence the human annotation of AI training data.

Sociocognitive Biases in Face Perception and Classification

Face classification systems rely on human judgments of facial attributes, such as race, gender, and emotion expression—judgments that can be influenced by a perceiver's social identity as well as stereotypes and prejudiced attitudes [21,23]. As a result, face annotation represents a critical entry point for the cultural transmission of inequality into AI systems [24]. Social psychology has documented several ways in which intergroup bias shapes the perception and classification of faces.

Cross-race effect. People are more likely to confuse or misidentify faces from other racial groups—a robust phenomenon driven by perceptual expertise and motivation to individuate ingroup members [25,26]. In face-recognition datasets, the **cross-race effect** can reduce labeling accuracy for minority faces, leading to poorer identification performance.

Hypodescent bias. Humans tend to classify multiracial faces as belonging to their lower-status groups, particularly when forced to assign a single racial category—an effect known as **perceptual hypodescent** [27–29]. The bias is stronger among individuals endorsing right-wing ideologies, zero-sum beliefs, and racial essentialism [30–32]. Such labeling biases can systematically distort AI classification thresholds, consistent with hypodescent-like patterns observed in image-to-text models [33].

Gendered race bias. Racial stereotypes also influence gender classification. For example, Black faces are more likely to be perceived as masculine, whereas East Asian faces are perceived as

more feminine, a **gendered-race effect** is consistent with White American stereotypes of Black people as aggressive and Asian people as submissive [34]. This labeling bias can introduce racial distortions into gender classification datasets.

Stereotyped Emotion Perception. Emotion judgments are systematically shaped by racial and gender stereotypes: White Americans more readily perceive anger in male and Black or Arab faces, and sadness or fear in female and Asian faces [35–38]. Consequently, annotators may systematically mislabel expressions, causing AI systems to learn stereotype-consistent mappings between social groups and emotions.

Perceptual Dehumanization. Race and gender even influence judgments of a person's humanness [39,40]. This bias emerges early in face perception, with ingroup faces encoded more readily than outgroup faces [41–43]. As a result, annotators may be less likely to label women and racial minorities as "people." An extreme example occurred when Google Photos labeled images of Black people as "gorillas" [44]. More subtle forms of dehumanization have been observed in search engines and language models, which disproportionately associate the category *person* with men rather than women [5,45].

Intergroup bias in judgments of harm

Prejudices, stereotypes, and ideologies shape judgments of harm, such that harmful acts are judged more severely when committed by racial minority group members [46–48]. This asymmetry reflects greater empathy for ingroup members and a tendency to attribute

outgroup transgressions to stable, immoral dispositions [49–51]. These biases can influence the annotation of social media used to train content moderation systems: posts authored by Black or Arab individuals are more likely to be labeled as offensive or hateful, particularly by annotators with conservative ideologies [52], and models trained on these labels reproduce these biases [53]. Similar distortions arise when LLMs are used as annotators, reflecting stereotypes embedded in their training corpora [54,55].

Stereotyping and Dehumanization in Health Care

Historical racial disparities in U.S. health care reflect, in part, prejudice, stereotypes, and dehumanization that lead physicians to spend less time with minority patients, underestimate their pain, and provide lower-quality care [56–60]. Medical AI systems trained on data reflecting biased clinical decisions and treatment disparities will reproduce and perpetuate these disparities [61–67]. For example, an AI health-care allocation system trained on historical data in which Black patients were undertreated significantly underpredicted the care they required, reinforcing existing disparities [68].

Stereotyping in Employment Decisions

Gender and racial disparities in employment are driven in part by stereotypes portraying women and minorities as less qualified for workplace roles [69,70]. Even with identical résumés, women and ethnic minorities are less likely to be interviewed, evaluated positively, or hired [71,72], and stereotypes portraying them as less agentic and competent create barriers to promotion and leadership [73,74]. These patterns are learned and replicated by AI systems

trained on a company's hiring history [75,76]—an effect famously illustrated by Amazon's résumé-screening tool, which penalized female applicants because it was trained on the company's historically male-dominated hiring patterns [8].

Effects of Mental States, Instructions, and Incentives

Annotation quality is also shaped by situational factors. Fatigue, cognitive load, and low accountability increase reliance on stereotypes [77,78], while task instructions and incentive structures influence how annotators interpret tasks and who chooses to perform them [79–82]. These factors, which can systematically enhance bias in labeling behavior, are rarely considered in dataset construction, yet their risks are exacerbated by a reliance on annotators in developing countries who often work under conditions of high pressure, precarity, and powerlessness [83].

Social-Cognitive Biases in Language

Human language is imbued with stereotypes, prejudices, culture, and status—in vocabulary, word choice, narrative structure, grammar, and syntax [84–86]. Consequently, LLMs trained on human text corpora reproduce these patterns [87,88]. These biases arise through three broad features of language: disparities in who produces text, disparities in who and what is discussed, and systematic associations between social groups and attributes.

Author Disparities

Although LLMs are increasingly trained on Internet text, their training corpora include books and newspapers with longstanding representational disparities. Historically, women and minority groups have been underrepresented as authors, and many widely used corpora (e.g., the *Wall Street Journal*) remain dominated by male authors writing for older, high-SES, politically conservative audiences [89]. Models trained on such data perform less well on language produced by younger and more diverse populations [90]. Moreover, minority groups are often represented through the lens of majority-group authors rather than by minority authors themselves, further skewing group portrayals [91].

Language data also reflect the overrepresentation of Western, Educated, Industrialized, Rich, and Democratic (WEIRD) populations [92]. Although multilingual resources continue to expand [93,94], languages used by “non-WEIRD” populations remain underrepresented and are served by lower-quality models [95,96]. As LLMs are increasingly trained on AI-generated text, these disparities may become further entrenched, potentially leading to cultural homogenization or "model collapse" [97].

Content Disparities

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 [98]. Consequently,

White, affluent men are more likely to be treated as the default category of “person” in the texts used to train LLMs [45,99].

These disparities are especially pronounced in social media, where political, moralized, emotional, and ingroup-focused content is overrepresented [100]. Engagement-optimizing algorithms can further exacerbate these patterns by promoting such content and shaping users' perceptions of what is normative, creating a self-reinforcing cycle of biased content generation.

Group–Attribute Associations

Linguistic biases also arise from systematic differences in how social groups are described in training text. Human language disproportionately associates dominant groups with positive attributes such as competence and leadership, while linking marginalized groups to negative or subordinate attributes. Gender is especially salient because it is explicitly encoded in many languages (e.g., pronouns and grammatical gender) [101,102]. As a result, gender stereotypes—such as associations of women with domestic roles and men with leadership or technical roles—are reproduced in word embeddings and LLMs [12,103]. These associations often exaggerate contemporary labor-market realities: occupations such as economist, pharmacist, and manager are now approximately 50% women yet remain male-stereotyped in LLMs because of persistent associations between men and work [103].

Comparable patterns appear for race, immigration status, social class, and health-related stigmas, with marginalized groups more strongly associated with poverty, criminality,

animalistic language, and negative affect [55,98,104,105]. Although the specific stereotypes attached to these groups change over time, their negative valence has remained remarkably stable [106]. Thus, biased group–attribute associations are a pervasive feature of human language and a persistent source of bias in language-based AI systems.

Biases in Training Data Selection

The choice of training data is itself a human decision, guided in part by developers' awareness of the social consequences of their systems. Even without intending to introduce bias, developers may select datasets that systematically underrepresent or exclude vulnerable populations, producing models that disproportionately harm those groups.

The tendency to recognize bias in others while overlooking it in oneself is known as the **bias blind spot** [107]. This pervasive bias reflects the tendency to judge oneself through introspection but others through their observable behavior [108]. In the context of prejudice, members of socially dominant groups, such as men or White majority group members, are less likely to notice a lack of diversity [109,110] or acknowledge inequality and discrimination [111]. As a result, developers from socially-dominant groups may unintentionally select training data that lack diversity or contain systematic biases (**Box 1**).

Blind spots in image data selection

Historically, face-image datasets have been heavily skewed toward White faces, leading models to perform worse for individuals with darker skin, especially women of color [112,113]. By

contrast, demographically balanced datasets such as FairFace [113] and DiveFace [114] improve accuracy across race, gender, and age while reducing performance disparities.

Blind spots in language data selection

The composition of language corpora likewise shapes LLM behavior. Although blogs, forums, and social media broaden demographic representation, they also disproportionately contain political, moralized, emotional, and polarized content [115,116]. When such datasets are selected to train an AI system, models may overestimate the prevalence, normativity, or acceptability of extreme or divisive content, including overt racism and sexism.

Proxy effects

Bias blind spots can also contribute to **proxy effects**—hidden correlations between protected social categories (e.g., race, gender, or age) and other variables that produce discriminatory outcomes despite efforts to remove demographic information from training data [117]. For example, zip code, occupation, education, and consumer behavior often serve as proxies for race, gender, or socioeconomic status because they systematically covary with these characteristics [118]. Problems discovered with the COMPAS recidivism tool illustrated this dynamic: although race was excluded from the training data, geographic variables correlated with race reproduced racial disparities in risk scores [119]. Because proxy effects are difficult to detect, bias blind spots can further obscure their role, allowing discriminatory outcomes to persist despite apparently unbiased models (see **Box 1**).

Temporal effects

Training data selection is also shaped by *when* data were produced. Because language, norms, and social attitudes change over time, older texts often include stronger stereotypes than more contemporary texts [55,90,120]. Consequently, LLMs trained on decades-old books and news articles are likely to inherit historical prejudices. These temporal effects can be amplified by the pursuit of scale: expanding training datasets with older sources may inadvertently increase social bias. Like proxy effects, temporal effects arise from structural properties of the data that a developer's bias blind spot may fail to recognize.

Multimodal Compounding of Biases

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 text, images, numerical data, and human-generated judgments [121]. 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 [122].

Furthermore, multiplicative statistical effects during algorithmic tuning can exacerbate the computational amplification of bias [123,124].

Multimodal compounding also obscures the human origins of algorithmic bias. Errors may appear to arise from complex model interactions rather than identifiable human inputs, diffusing accountability and contributing to **algorithmic bias laundering** (**Box 2**). Understanding how human biases accumulate across modalities will be essential for diagnosing algorithmic bias and designing effective interventions as such models become increasingly complex.

Section summary

Human prejudices are transmitted to AI systems through multiple psychological mechanisms that shape data annotation, language production, and other human behaviors involved in creating training data. Although these mechanisms often interact in complex ways that obscure their effects, identifying them provides a foundation for understanding, diagnosing, and ultimately reducing algorithmic bias.

Table 1. Human Sources of Bias in AI Creation

Stage of AI Creation	Psychological mechanism	Illustrative examples	Algorithmic consequence
Data Annotation	Social perception and categorization biases	Cross-race effect; perceptual hypodescent; gendered race effect; stereotyped emotion perception; perceptual dehumanization	Subjective labels encode stereotypes, reducing performance and increasing errors for marginalized groups
	Intergroup judgment biases	Differential judgments of harm and offensiveness; ideology-dependent content moderation	Models learn biased standards for toxicity, threat, and harmfulness
	Institutional decision biases	Historical discrimination in healthcare and employment	AI reproduces disparities in diagnosis, treatment, hiring, and promotion
	Annotator context and motivation	Fatigue, cognitive load, task instructions, incentives, annotator demographics	Situational factors increase stereotype-based labeling
Language Data Curation	Author disparities	Texts underrepresent women and minorities, overrepresent WEIRD and English-language populations	Models reproduce dominant cultural views; generalize poorly across populations
	Content disparities	Overrepresentation of political, moralized, emotional, and high-status group content	Models treat dominant groups and content as more normative or representative
	Group-attribute associations	Men associated with high-power roles; marginalized groups with negative attributes	Models reproduce and amplify social stereotypes
Training Data Selection	Bias blind spots and dataset composition	Underrepresentation of minority groups; WEIRD and English-language dominance	Models generalize poorly across demographic groups and cultures
	Structural biases in training data	Content frequency disparities; group–attribute associations; proxy variables; temporal effects	AI reproduces historical stereotypes and hidden structural inequalities

Human Sources of Bias in AI Consumption

The second major entry point for human sources of algorithmic bias occurs during the consumption of AI outputs: how users perceive, interpret, and act on algorithmic recommendations. These responses are shaped by well-established processes in social cognition, motivation, and decision making that influence whether AI outputs are trusted, scrutinized, or rejected. They also generate behavioral feedback—for example, through clicks, accepted recommendations, and interactions with generated content—that is incorporated into subsequent model training and optimization [18].

Overreliance and Trust in AI

A key determinant of whether biased AI outputs translate into real-world discrimination is the degree to which they are trusted and accepted by users. When AI systems are perceived as fair, objective, or technically superior, users are more likely to accept their recommendations uncritically, allowing biased outputs to influence consequential decisions [125].

Trust varies systematically across individuals. Users with less knowledge or experience with AI show greater trust in algorithmic outputs [126,127]. Trust also differs by age, with children and older adults relying more on internet search results than younger adults [128], and by gender, with men reporting greater trust in AI than women [129]. In turn, greater trust predicts stronger reliance on AI for decision making [130].

The influence of trust is compounded by AI's *veneer of objectivity*—the tendency to perceive computational outputs as inherently impartial [131]. This effect depends on context, however, as users are more skeptical of AI when it mimics uniquely human capacities such as moral reasoning [132,133].

Trust effects are further reinforced by **cognitive offloading**—the tendency to delegate cognitive effort to external tools. Consistent with research on automation bias, people reduce independent monitoring once an automated system is perceived as reliable, making them less likely to detect errors or question recommendations [134]. Likewise, greater reliance on AI decision tools is associated with lower critical thinking, particularly among younger and less-educated users [135].

Finally, disparities in trust and AI adoption can themselves generate inequality. Differences in access to, familiarity with, or willingness to use AI can produce unequal benefits from AI-assisted decision making, potentially reinforcing existing social inequalities [136,137].

Motivated Reasoning and Confirmation Bias

Even when users trust AI, they do not trust all outputs equally. Human interactions with AI are shaped by **motivated reasoning**—the tendency to process information in ways that serve one's personal or group-based goals [138,139]. Consistent with this idea, trust in AI depends partly on users' existing goals and beliefs [125,140,141], making users more likely to accept algorithmic outputs that support their existing views [142,143].

This tendency reflects **confirmation bias**, whereby people seek and more readily accept information supporting their existing beliefs [117,144]. Thus, users are more likely to trust and implement algorithmic outputs that align with their beliefs, while discounting or rejecting outputs that conflict with them [140–142]. The confident, authoritative tone of many generative AI systems increases the perceived credibility of belief-consistent responses [145] and AI chatbots that give sycophantic responses inflate a users' certainty about their existing attitudes [146], while bias blind spots may prevent users from recognizing these influences [107].

People generally prefer AI systems over humans when their outputs confirm their expectations [143]. Although users will often avoid algorithms after observing errors [147], this *algorithmic aversion* diminishes when users retain even minimal control over recommendations [148]. Thus, confirmation bias shapes people's preferences for both algorithms and their outputs.

Cognitive Heuristics, Ranking Effects, and Social Norms

Even in the absence of motivated reasoning, cognitive heuristics shape how AI outputs influence judgment. The **availability heuristic** causes people to rely disproportionately on information that is salient or easily retrieved, whereas **anchoring** causes early information to exert disproportionate influence on subsequent judgments [149]. Accordingly, users tend to rely more on top-ranked search results, default recommendations, and the first response generated by a language model [117,150].

Because AI systems prioritize and rank information rather than presenting it neutrally, they can exploit availability and anchoring heuristics to shape users' judgments. Recommender algorithms often exhibit **popularity bias**, disproportionately promoting already popular items and reinforcing existing exposure inequalities [151]. Experimental evidence shows that stereotype-consistent ranking effects influence social judgments: increasing the proportion of male exemplars in early search results strengthened beliefs that men were more suitable for particular occupations [5].

Rankings also communicate descriptive social norms. Popularity indicators such as *most viewed*, *trending*, or highly ranked recommendations signal what other people appear to value or endorse, creating a **conformity bias** [152]. Decades of social psychology show that individuals often conform to perceived descriptive norms, particularly under uncertainty [153,154]. Thus, algorithmic rankings may reinforce prejudice through two complementary mechanisms: by increasing the cognitive accessibility of biased information and by signaling that such information reflects prevailing norms.

Individual Differences in Prejudice and Ideology

Up to this point, we have described cognitive biases that can lead well-intentioned users to implement discriminatory AI outputs. However, individual differences in prejudice and ideology can also directly shape AI use. Research shows that people with prejudiced attitudes and rightwing ideologies are more likely to endorse stereotypes, discriminate against outgroups,

and justify existing social hierarchies [139,155]. Although overt expressions of prejudice have decreased since the mid-19th century due to changing norms, blatant racism and sexism still persist and continue to drive discrimination [156–158].

Explicit prejudice can influence AI use in at least two ways. First, prejudiced users may deliberately seek out and implement AI outputs that support discriminatory goals or confirm existing stereotypes while discounting egalitarian recommendations [159,160]. Second, AI systems can provide justification and plausible deniability for discriminatory decisions. Rather than expressing prejudice directly, users could attribute unfair outcomes to an ostensibly objective algorithm—an intentional form of *bias laundering* (see Box 2).

Although these predictions require further empirical testing, they suggest that individual differences in prejudice and ideology shape not only whether biased AI outputs are accepted, but also whether they are used to maintain and legitimize existing social inequalities. Within the Human–AI Loop, prejudice serves as a motivational force that perpetuates bias between AI systems, individuals, and society.

Section summary

The effects of AI depend not only on algorithms but also on the psychology of their users. Trust, motivated reasoning, cognitive heuristics, and prejudice determine how algorithmic outputs are interpreted and acted upon. Because these responses also generate behavioral data that

inform future model training, these psychological processes both implement and reinforce algorithmic bias through the Human–AI Loop.

Table 2. Human Sources of Bias in the Use of AI Systems

Sources of Bias	Psychological mechanisms	Effects on judgment and behavior
Trust and reliance on AI	Overreliance and automation bias, Cognitive offloading, Veneer of objectivity, Individual differences in trust and AI adoption	Biased outputs are accepted with insufficient scrutiny, increasing implementation of discriminatory recommendations and unequal access to AI benefits
Motivated reasoning	Confirmation bias, Algorithmic aversion and selective reliance, Bias blind spots	Users preferentially accept stereotype-consistent outputs and discount contradictory information, reinforcing existing beliefs and feeding biased interactions back into AI systems
Heuristics and decision biases	Availability heuristic, Anchoring, Popularity bias, Conformity and Social norms	Ranked or popular outputs shape beliefs about what is true or normative, while prejudice facilitates discriminatory use and bias laundering, strengthening the Human–AI Loop
Individual differences in prejudice and ideology	Explicit and implicit prejudice, Social dominance orientation, System-justifying motivations	Explicit prejudice can lead users to selectively adopt discriminatory AI outputs and use algorithms to justify or legitimize unequal decisions

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

The human sources of bias we have described at the stages of AI creation and consumption are not independent; they reinforce one another over time. Human prejudices are embedded in training data, biased data shape model outputs, and model outputs influence human beliefs and behavior, generating new data that feed subsequent models. Over time, this feedback loop can entrench and amplify inequalities while producing qualitatively new forms of bias (Box 3).

Importantly, these effects extend beyond individual decisions to reshape broader social and cultural structures. As people increasingly rely on AI to select information, recommend opportunities, and guide consequential decisions, algorithmic outputs influence the distribution of resources, social interactions, and cultural representations. For example, biased hiring algorithms can alter workforce composition, recommender systems can amplify media exposure and political polarization, and search engines or generative AI can reinforce stereotypes by repeatedly presenting them as normative or accurate [5,18,116,124,161]. This full loop pattern has been demonstrated by Vlasceanu and Amodio in the context of gender inequality and stereotypes [5].

The result is an altered social reality that then creates the experiences and datasets from which future AI systems learn. That is, AI systems do not merely reflect existing prejudice—they reshape the social world in ways that become encoded into subsequent generations of AI. The result is a self-reinforcing Human–AI Loop in which psychological biases, algorithmic outputs, and societal structures continually coevolve, potentially producing new forms of discrimination.

Although our emphasis has been on human sources of algorithmic bias, these processes operate alongside effects introduced during model development. Even holding human inputs constant, technical choices—including model architecture, feature representations, objective functions, optimization procedures, hyperparameters, and evaluation criteria—shape what patterns models learn, which errors they prioritize, and what information is amplified or

ignored. Thus, algorithmic bias cannot be attributed solely to training data; it also reflects technical decisions that determine how human-generated information is represented and transformed [2,14,19,162,163].

The Human–AI Loop proposed here integrates three interacting levels of analysis: individual psychological processes that shape AI creation and consumption, computational processes that transform human biases into algorithmic outputs, and societal structures that are shaped by these outputs and, in turn, reinforce their biases by providing data to train future AI systems. By specifying psychological mechanisms operating at each stage of this cycle, our framework integrates individual, computational, and structural levels of analysis into a unified account of algorithm-mediated discrimination (see Box 3).

Breaking the Loop: Implications for bias reduction

Traditional human-in-the-loop approaches assume that human oversight improves AI systems [164]. The caveat, however, is that humans are themselves prone to bias, and thus human oversight is not inherently bias reducing. By identifying the psychological mechanisms through which human biases enter and amplify AI systems, the Human-AI Loop model clarifies where human interventions are most likely to be effective. Below, we outline intervention strategies for both AI creation and AI consumption, drawing on insights from social and cognitive psychology.

Interventions at Input: Reducing Bias in Model Creation

Interventions targeting AI creation focus on the social-cognitive processes underlying data generation and dataset selection.

Promoting fair and diverse annotation

Because many AI systems rely on human-labeled data, reducing bias in annotation is crucial. Recruiting demographically and culturally diverse annotators broadens the perspectives informing labels and reduces the dominance of majority viewpoints. Annotation protocols can incorporate bias-awareness prompts, clear task instructions, and structured practice with feedback—procedures shown to reduce reliance on heuristics and stereotypes [81,82]. Practical factors such as fatigue, time pressure, distraction, and incentives should also be managed, as these conditions increase reliance on cognitive shortcuts.

Alignment procedures that discourage models from assigning different labels solely because of group identity can further improve fairness when biases are well defined and measurable. However, they are less effective when bias depends on indirect, culturally specific, or context-dependent meanings that cannot be isolated through simple label or identity swaps [87]. 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 [165].

Reducing bias in dataset selection

Bias can also enter AI systems through the choice of training data. Choice of dataset 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. Moreover, proportional representation is often insufficient; achieving comparable performance across demographic groups frequently requires equal representation in training data, as demonstrated in face classification models [113].

For LLMs, curating training corpora to reduce prejudice, expanding multilingual resources, and incorporating underrepresented communication modalities (e.g., sign language) can improve demographic and cultural coverage during model alignment [166–168]. As with annotation, documenting dataset composition, geographic origin, annotator demographics, and historical context helps developers and users identify potential sources of bias [165]. Collectively, these practices shift dataset construction from maximizing scale alone toward prioritizing fairness.

Interventions at Output: Reducing Bias in AI Use

Although human prejudice may exert its strongest effects during AI creation, those input-stage biases are often difficult to detect and modify. Psychology-based interventions may therefore be especially effective at the output stage, where users interpret and implement AI recommendations through their judgments and behavior. Psychology is well positioned to inform interventions at the point of AI consumption, drawing on research on prejudice

reduction, judgment, and self-regulation, and evidence suggests a combination of both individual- and structural-level interventions are most effective [169,170].

Individual-level interventions

Because users often accept algorithmic recommendations without sufficient scrutiny, interventions should strengthen both the detection and correction of bias [171,172]. Bias detection could 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 [173,174]. Bias correction can be facilitated through strategies that engage deliberative responses and goal-driven action plans, such as implementation intentions (e.g., "If I encounter an AI recommendation involving a protected social group, then I will verify it using an independent source") [175–177].

Insights from judgment and decision-making research further suggest strategies to reduce overreliance on biased AI outputs. Encouraging users to consider alternative hypotheses, generate counterfactuals, or explicitly justify decisions can reduce anchoring, confirmation bias, and overreliance on salient or belief-consistent algorithmic outputs [150,178]. Interventions should also aim to calibrate, rather than simply reduce, trust in AI by helping users distinguish situations in which algorithmic recommendations are likely to be reliable from those requiring greater scrutiny [125,131].

Structural interventions

Structural interventions reduce bias by altering decision environments and policy rather than targeting individual-level attitudes and self-regulation [179,180]. Similar strategies can be applied to AI use. Interface features such as presenting multiple alternatives, displaying uncertainty estimates, restricting systems to specified decision contexts (i.e., to avoid inappropriate uses), or requiring verification against non-AI sources may reduce overreliance on single outputs [151,181].

At the policy level, transparency requirements, documentation standards, and algorithmic impact assessments can serve as structural guardrails that shape how AI systems are interpreted and applied [182]. Such measures not only constrain technology but raise awareness of how AI outputs are perceived, evaluated, and implemented. These measures can be complemented by technological tools such as automated bias detection systems [183] and personalized AI agents designed to steer users away from harmful outcomes [184].

Section Summary

From the perspective of the Human–AI Loop, algorithmic bias cannot be mitigated by changing algorithms alone. Because human psychological processes shape both the creation and consumption of AI systems, they should be viewed not only as a source of algorithmic bias but also as a primary target for intervention. Accordingly, effective bias mitigation requires coordinated psychological, computational, structural interventions across the AI lifecycle.

A Psychology of Algorithmic Bias

More broadly, our analysis highlights the central 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 structures, then algorithmic bias cannot be fully explained or mitigated without psychological science.

This perspective calls for closer collaboration between computer scientists, data scientists, and social and cognitive psychologists. Psychology offers mechanistic theories of perception, categorization, stereotyping, motivated reasoning, judgment, and decision making that explain how human biases enter AI systems during both their creation and consumption. These theories complement computational analyses by identifying the psychological mechanisms underlying algorithmic bias and informing interventions that target them.

Of course, not all forms of algorithmic bias arise directly from human behavior: statistical artifacts, nonrepresentative sampling, model misspecification, and biased evaluation can introduce disparities even in the absence of explicit human prejudice [2,87]. Yet these technical failures are rarely independent of human judgment. Decisions about data selection, model design, evaluation, and deployment are themselves shaped by human cognition. A psychological perspective therefore complements computational analyses by explaining how technical failures emerge from, interact with, and ultimately influence human cognition and behavior.

Advancing a psychology of algorithmic bias will require studying the full Human–AI Loop rather than isolated components. This includes understanding how psychological biases shape dataset construction, how people interpret and act on AI outputs, how these interactions generate new training data, and how AI systems reshape social norms and institutions over time. Addressing these questions will require closer collaboration across disciplines as well as training that integrates computational methods with theories of human cognition and social behavior.

Concluding remarks

Despite its computational expression, algorithmic bias is fundamentally rooted in human psychology. We introduced the Human–AI Loop model, which conceptualizes algorithmic bias as emerging from dynamic interactions between humans and AI computation. By identifying the psychological pathways through which human biases enter AI systems during both their creation and use, this model provides a more complete account of how algorithmic discrimination emerges and persists. Ultimately, achieving fairer AI will require treating algorithmic bias not only as a computational challenge but also as a psychological one.

Glossary

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

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

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

Anchoring: The tendency to rely on initially-encountered information to heavily in decision making.

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.

Availability heuristic: The tendency to judge the likelihood or importance of something based on how easily examples come to mind.

Bias blind spot: The tendency to recognize bias in others while overlooking it in oneself.

Cognitive offloading: The delegation of cognitive effort to external tools such as automated systems.

Confirmation bias: The tendency to seek out and more readily accept information supporting one's existing beliefs

Conformity bias: The tendency to adjust one's judgments or behavior to align with the views of other people or signaled by algorithmic outputs.

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

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.

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

Motivated reasoning: the tendency to process information in ways that serve one's beliefs, goals, or desired conclusions.

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

Popularity bias: The tendency to favor information, products, or choices that are already popular, such as in recommender algorithms.

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

Proxy effects: Hidden correlations between protected social categories and other variables that produce discriminatory outcomes despite efforts to remove demographic information from training data.

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

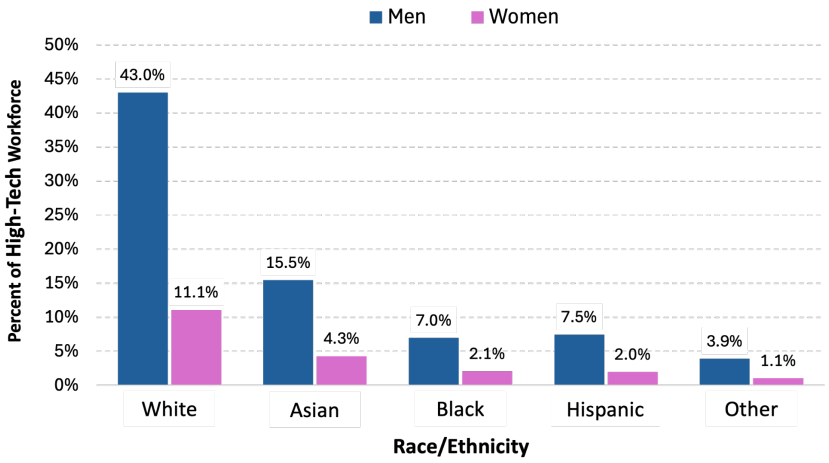
Stereotypes: Cognitive beliefs or expectations about the characteristics, traits, or behaviors of members of a social group.

Box 1: Organizational Effects on Algorithmic Bias

Algorithmic bias is shaped not only by the psychology of individual developers and users but also by the organizational cultures in which AI systems are created. Organizations determine who builds AI, whose perspectives are represented, which performance metrics are prioritized, and how competing goals such as accuracy, efficiency, fairness, and profitability are balanced.

Lack of diversity in the technology sector

Technology companies—particularly in the United States—remain demographically homogeneous, with leadership positions disproportionately occupied by White men [185]. Women, Black, Latino, disabled, and LGBTQ+ individuals are underrepresented, particularly in decision-making roles, as are workers from relatively low socioeconomic backgrounds.



Box 1, Figure 1. Race and gender composition of the US high-tech workforce in 2022, illustrating the under-representation of women and racial/ethnic minorities [185].

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 [69,186,187]. These blind spots shape which concerns are raised, harms anticipated, and design decisions prioritized. Highly publicized failures, such as the "racist soap dispensers" that failed to detect dark skin [188], illustrate these risks. Conversely, diverse teams are more likely to identify assumptions and potential harms that homogeneous groups overlook because members bring different experiences and perspectives to the evaluation of AI systems [189].

Box 1, Fig 2. The “racist soap dispenser” meme, illustrating the risk of technological bias blind spots, based on an incident where a soap dispenser detected the hand of a White person, but not of his Black colleague. However, when the Black person held a white paper towel, the dispenser worked.

Incentives and market pressures

Organizational culture interacts with market pressures to further exacerbate bias toward historically disadvantaged groups. AI systems are often optimized for engagement, growth, or efficiency rather than fairness or broader social impact [100]. In the absence of regulatory

constraints, bias mitigation may be deprioritized, particularly when harms affect groups with limited economic or political power.

Organizational power and accountability

Organizational power dynamics also shape algorithmic bias through relationships with annotators, contractors, and users. Annotators—often low-paid and geographically distant—have limited power to raise ethical concerns [83,190], increasing the likelihood that problematic assumptions remain embedded in AI systems. The emergence of AI auditing reflects growing recognition that organizational structures are critical for mitigating algorithmic bias [191].

Changing organizational cultures

Reducing algorithmic bias requires organizational cultures that value fairness alongside accuracy, efficiency, and profitability. This includes increasing demographic and disciplinary diversity, incentivizing the identification of potential harms, empowering employees and external auditors to raise concerns, and embedding fairness into organizational evaluation [192]. Regulatory frameworks such as the EU AI Act can reinforce these changes by aligning organizational incentives with societal goals.

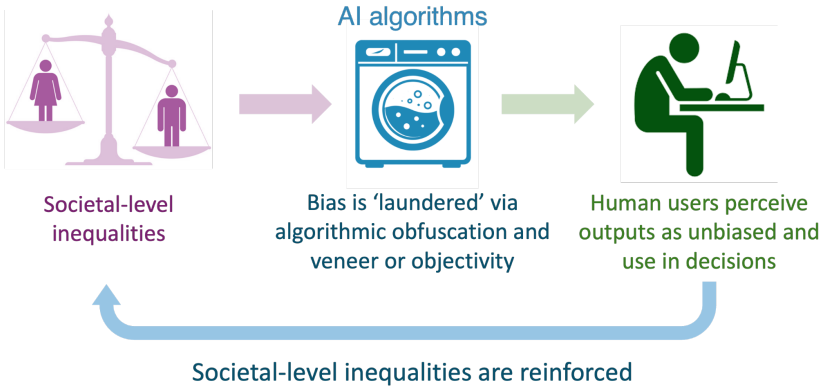
Box 2. Algorithmic Implicit Bias & Bias Laundering

According to the Human-AI Loop model, algorithmic bias is not merely an artifact of computation, but a novel expression of human prejudice via the mediating interface of AI. This perspective reimagines the construct of implicit bias: whereas classic theories conceptualize it as an intra-individual phenomenon—the indirect expression of a person's prejudices or stereotypes in their behavior [172,193,194]—algorithmic implicit bias reflects a multi-agent phenomenon in which prejudice is transmitted between independent human and computational entities.

Research using the Word Embedding Association Test and queries to AI chatbots demonstrates the first stage of this process: transmission of human social biases embedded in language to AI representations [87,88]. Other work, which shows that biased AI outputs alter users' stereotypes and social judgments, demonstrates the second stage: transmission from AI back to human cognition [5,161]. The full-loop expression of algorithmic implicit bias occurs when prejudice passes from one set of humans to AI and, unwittingly, to another human who is unaware of its origin [5].

This conceptualization sidesteps longstanding debates over whether implicit bias involves unconscious or unintentional mental processes [195,196]. Algorithmic implicit bias is defined not by any single actor's mental state but by the transmission of prejudice across human and computational agents, regardless of whether any agent intended or recognized the discriminatory outcome. For these reasons, algorithmic implicit bias is especially insidious:

because prejudice becomes obscured by AI's opacity and veneer of objectivity [106,131], its origins in human prejudice are easily masked—an effect we term *algorithmic bias laundering* (Box 1 Fig 1).



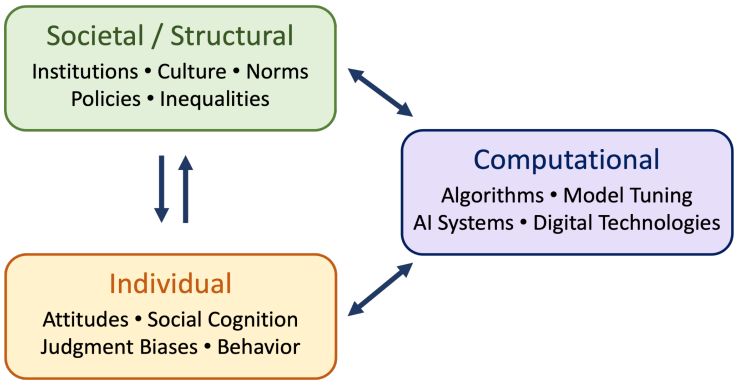
Box 2, Figure 1. Algorithmic bias laundering: Effects of human prejudice on AI systems are obscured by the algorithmic process and presented to end-users as bias-free.

Although algorithmic bias laundering often occurs inadvertently, it can also be exploited intentionally. Developers, organizations, or end users may knowingly deploy biased training data or invoke biased outputs to advance discriminatory goals while attributing the results to an ostensibly objective AI system. Whether inadvertent or intentional, algorithmic bias laundering obscures responsibility and makes discrimination harder to detect and correct.

This perspective extends the concept of implicit bias from a property of individual cognition to one involving sociotechnical systems. In doing so, it provides a psychological framework for understanding how prejudice is transmitted, obscured, and perpetuated through interactions among humans, AI, and social institutions.

Box 3. A Computational Level of Analysis for Theories of Prejudice

Our analysis suggests that theories of prejudice require an additional level of analysis for the digital age. Traditionally, explanations have focused on the individual level—emphasized in psychology—or the societal level—emphasized in sociology and political science. The Human–AI Loop model identifies AI systems as a third, computational level of analysis that operates between individuals and social structures. This computational level provides a distinct mechanism for representing and expressing prejudice which interacts dynamically with both individual- and societal-level biases.



Box 3, Figure 1. Algorithmic systems constitute a distinct computational level of analysis linking individual cognition and societal structures. Rather than simply reflecting human prejudice, AI systems transform and transmit social biases between individuals and institutions through recursive feedback loops.

Although the computational processes through which AI systems learn, represent, and express social bias are distinct from those operating at the individual or societal levels, they share some general features. Like individuals, AI systems learn from experience—in this case, large collections of human behavior, language, and judgments—and express these learned associations in interactions with users. Like societal structures, AI systems encode statistical

regularities aggregated across collectives rather than the beliefs of any single individual, akin to culturally shared stereotypes that persist independent of individual-level beliefs. Also like societal processes, due to the perception of algorithmic decisions as objective or institutionally sanctioned, AI systems can shift responsibility for discriminatory outcomes from individuals to opaque technological infrastructures, legitimizing and stabilizing existing inequalities.

By bridging levels of analysis, as described by the Human-AI Loop model, AI systems provide a new pathway through which prejudice is transmitted between individual minds and social structures. By formalizing social biases into computational representations and reinserting them into everyday decisions, they link cognition, culture, and institutions in a recursive feedback loop. This computational level expands theories of prejudice beyond traditional individual–societal frameworks, positioning AI not merely as a reflection of existing bias but as an active component in its generation, maintenance, and amplification.

References

1. Gomez, C. *et al.* (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
2. Hooker, S. (2021) Moving beyond "algorithmic bias is a data problem." *Patterns* 2
3. Noble, S.U. (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York University Press.
4. O'Neil, C. (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown Publishing Group.
5. 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
6. Zou, J. & Schiebinger, L. (2018) AI can be sexist and racist — it's time to make it fair. *Nature* 559, 324–326
7. Angwin, J. Larson, J., Mattu, S., & Kirchner, L. (2016) What Algorithmic Injustice Looks Like in Real Life. *ProPublica*. [Online]. Available: https://www.propublica.org/article/what-algorithmic-injustice-looks-like-in-real-life. [Accessed: 28-Aug-2025]
8. Dastin, J. (2018) Insight - Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/
9. Dodd, V. (2018) UK police use of facial recognition technology a failure, says report. *The Guardian*. https://www.theguardian.com/uk-news/2018/may/15/uk-police-use-of-facial-recognition-technology-failure
10. Hill, K. (2020) Wrongfully Accused by an Algorithm. *The New York Times*. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html
11. Tay, L. *et al.* (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
12. Bolukbasi, T. *et al.* (2016) Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. in *Advances in Neural Information Processing Systems*, 29
13. Favaretto, M. *et al.* (2019) Big Data and discrimination: perils, promises and solutions. A systematic review. *J Big Data* 6, 12
14. Hardt, M. *et al.* (2016) Equality of Opportunity in Supervised Learning. in *Advances in Neural Information Processing Systems*, 29
15. Lum, K. and Isaac, W. (2016) To Predict and Serve? *Significance* 13, 14–19
16. Barocas, S. *et al.* (2023) *Fairness and Machine Learning: Limitations and Opportunities*, MIT Press
17. Ensign, D. *et al.* (2018) Runaway Feedback Loops in Predictive Policing. in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pp. 160–171
18. Mansoury, M. *et al.* (2020) Feedback Loop and Bias Amplification in Recommender Systems. in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 2145–2148
19. Selbst, A.D. *et al.* (2019) Fairness and Abstraction in Sociotechnical Systems. in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 59–68

20. Fiske, S.T. (1998) Stereotyping, prejudice, and discrimination. In *The handbook of social psychology, Vols. 1-2, 4th ed*, pp. 357–411, McGraw-Hill
21. Kawakami, K. *et al.* (2017) Intergroup perception and cognition: An integrative framework for understanding the causes and consequences of social categorization. In *Advances in experimental social psychology* 55, pp. 1–80, Elsevier
22. Williams, N.W. *et al.* (2025) When Conservatives See Red but Liberals Feel Blue: Labeler Characteristics and Variation in Content Annotation. *The Journal of Politics* DOI: 10.1086/735397
23. Chen, J.M. (2019) An integrative review of impression formation processes for multiracial individuals. *Social & Personality Psych* 13, e12430
24. Chen, Y. and Joo, J. (2021) Understanding and Mitigating Annotation Bias in Facial Expression Recognition. presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14980–14991
25. Hugenberg, K. *et al.* (2010) The categorization-individuation model: An integrative account of the other-race recognition deficit. *Psychological Review* 117, 1168–1187
26. 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
27. Chen, J.M. and Hamilton, D.L. (2012) Natural ambiguities: Racial categorization of multiracial individuals. *Journal of Experimental Social Psychology* 48, 152–164
28. Ho, A.K. *et al.* (2017) "You're one of us": Black Americans' use of hypodescent and its association with egalitarianism. *Journal of Personality and Social Psychology* 113, 753–768
29. Peery, D. and Bodenhausen, G.V. (2008) Black + white = black: hypodescent in reflexive categorization of racially ambiguous faces. *Psychol Sci* 19, 973–977
30. Ho, A.K. *et al.* (2015) Essentialism and Racial Bias Jointly Contribute to the Categorization of Multiracial Individuals. *Psychol Sci* 26, 1639–1645
31. Krosch, A.R. *et al.* (2013) On the ideology of hypodescent: Political conservatism predicts categorization of racially ambiguous faces as Black. *Journal of Experimental Social Psychology* 49, 1196–1203
32. 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
33. Wolfe, R. *et al.* (2022) Evidence for Hypodescent in Visual Semantic AI. in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1293–1304
34. Johnson, K.L. *et al.* (2012) Race is gendered: How covarying phenotypes and stereotypes bias sex categorization. *Journal of Personality and Social Psychology* 102, 116–131
35. Bijlstra, G. *et al.* (2010) The social face of emotion recognition: Evaluations versus stereotypes. *Journal of Experimental Social Psychology* 46, 657–663
36. Brooks, J.A. *et al.* (2018) Stereotypes Bias Visual Prototypes for Sex and Emotion Categories. *Social Cognition* 36, 481–493
37. Halberstadt, A.G. *et al.* (2022) Racialized emotion recognition accuracy and anger bias of children's faces. *Emotion* 22, 403–417
38. Plant, E.A. *et al.* (2000) The Gender Stereotyping of Emotions. *Psychology of Women Quarterly* 24, 81–92
39. Goff, P.A. *et al.* (2008) Not yet human: Implicit knowledge, historical dehumanization, and contemporary consequences. *Journal of Personality and Social Psychology* 94, 292–306

40. Haslam, N. and Loughnan, S. (2014) Dehumanization and Infrahumanization. *Annual Review of Psychology* 65, 399–423
41. Ito, T.A. and Urland, G.R. (2005) The influence of processing objectives on the perception of faces: An ERP study of race and gender perception. *Cognitive, Affective, & Behavioral Neuroscience* 5, 21–36
42. Krosch, A.R. and Amodio, D.M. (2019) Scarcity disrupts the neural encoding of Black faces: A socioperceptual pathway to discrimination. *Journal of personality and social psychology* 117, 859
43. Ratner, K.G. and Amodio, D.M. (2013) Seeing "us vs. them": Minimal group effects on the neural encoding of faces. *Journal of experimental social psychology* 49, 298–301
44. Simonite, T. (2018) When It Comes to Gorillas, Google Photos Remains Blind*Wired*
45. Bailey, A.H. *et al.* (2022) Based on billions of words on the internet, people = men. *Science Advances* 8, eabm2463
46. Casey, J.P. *et al.* (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
47. 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
48. 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
49. Brewer, M.B. (1999) The Psychology of Prejudice: Ingroup Love and Outgroup Hate? *Journal of Social Issues* 55, 429–444
50. Bruneau, E.G. *et al.* (2017) Parochial Empathy Predicts Reduced Altruism and the Endorsement of Passive Harm. *Social Psychological and Personality Science* 8, 934–942
51. Pettigrew, T.F. (1979) The Ultimate Attribution Error: Extending Allport's Cognitive Analysis of Prejudice. *Pers Soc Psychol Bull* 5, 461–476
52. Sap, M. *et al.* (2019) The Risk of Racial Bias in Hate Speech Detection. in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1668–1678
53. Davani, A.M. *et al.* (2023) Hate Speech Classifiers Learn Normative Social Stereotypes. *Transactions of the Association for Computational Linguistics* 11, 300–319
54. Das, A. *et al.* (2024) Investigating Annotator Bias in Large Language Models for Hate Speech Detection. ArXiv
55. Garg, N. *et al.* (2018) Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, E3635–E3644
56. Green, A.R. *et al.* (2007) Implicit Bias among Physicians and its Prediction of Thrombolysis Decisions for Black and White Patients. *J GEN INTERN MED* 22, 1231–1238
57. Hoffman, K.M. *et al.* (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
58. Johnson, R.L. *et al.* (2004) Patient Race/Ethnicity and Quality of Patient–Physician Communication During Medical Visits. *Am J Public Health* 94, 2084–2090
59. Penner, L.A. *et al.* (2016) The Effects of Oncologist Implicit Racial Bias in Racially Discordant Oncology Interactions. *JCO* 34, 2874–2880

60. Zavala, V.A. *et al.* (2021) Cancer health disparities in racial/ethnic minorities in the United States. *Br J Cancer* 124, 315–332
61. Celi, L.A. *et al.* (2022) Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. *PLOS Digital Health* 1, e0000022
62. Chen, R.J. *et al.* (2023) Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat. Biomed. Eng* 7, 719–742
63. Larrazabal, A.J. *et al.* (2020) Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences* 117, 12592–12594
64. Markowitz, D.M. (2022) Gender and ethnicity bias in medicine: a text analysis of 1.8 million critical care records. *PNAS Nexus* 1
65. McKinney, S.M. *et al.* (2020) International evaluation of an AI system for breast cancer screening. *Nature* 577, 89–94
66. Norori, N. *et al.* (2021) Addressing bias in big data and AI for health care: A call for open science. *Patterns* 2
67. Seyyed-Kalantari, L. *et al.* (2021) Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med* 27, 2176–2182
68. Obermeyer, Z. *et al.* (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 447–453
69. Heilman, M.E. (2012) Gender stereotypes and workplace bias. *Research in Organizational Behavior* 32, 113–135
70. Quillian, L. *et al.* (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
71. Steinpreis, R.E. *et al.* (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
72. 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
73. Kray, L.J. *et al.* (2026) Psychological drivers of gender disparities in leadership paths. *Trends in Cognitive Sciences* 30, 515–529
74. Rosette, A.S. *et al.* (2008) The White standard: Racial bias in leader categorization. *Journal of Applied Psychology* 93, 758–777
75. Chen, Z. (2023) Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanit Soc Sci Commun* 10, 1–12
76. 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
77. 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
78. Paolini, S. *et al.* (2009) Accountability moderates member-to-group generalization: Testing a dual process model of stereotype change. *Journal of Experimental Social Psychology* 45, 676–685

79. Ding, Y. *et al.* (2022) Impact of Annotator Demographics on Sentiment Dataset Labeling. *Proc. ACM Hum.-Comput. Interact.* 6, 519:1-519:22
80. 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*, pp. 823–835
81. McDaniel, M.A. *et al.* (2007) Situational Judgment Tests, Response Instructions, and Validity: A Meta-Analysis. *Personnel Psychology* 60, 63–91
82. Parmar, M. *et al.* (2023) Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions. in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1779–1789
83. Le Ludec, C. *et al.* (2023) The problem with annotation. Human labour and outsourcing between France and Madagascar. *Big Data & Society* 10, 20539517231188723
84. Kashima, Y. *et al.* (2008) *Stereotype Dynamics: Language-based Approaches to the Formation, Maintenance, and Transformation of Stereotypes*, Taylor & Francis
85. Maass, A. *et al.* (1989) Language use in intergroup contexts: The linguistic intergroup bias. *Journal of Personality and Social Psychology* 57, 981–993
86. Rhodes, M. *et al.* (2025) How generic language shapes the development of social thought. *Trends in Cognitive Sciences* 29, 122–132
87. Bai, X. *et al.* (2025) Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences* 122, e2416228122
88. Caliskan, A. *et al.* (2017) Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 183–186
89. Waldfogel, J. (2025) The Welfare Effects of Gender-Inclusive Intellectual Property Creation: Evidence from Books. *Journal of Political Economy* 133, 2229–2264
90. 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)*, pp. 483–488
91. Wang, A. *et al.* (2025) Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nat Mach Intell* 7, 400–411
92. Henrich, J. *et al.* (2010) The weirdest people in the world? *Behavioral and Brain Sciences* 33, 61–83
93. Date, S. *et al.* (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
94. Grave, E. *et al.* (2018) Learning Word Vectors for 157 Languages. in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*
95. Jackson, J.C. *et al.* (2026) Large AI Models Have a Prioritization Problem: Policy Implications and Solutions. *Policy Insights from the Behavioral and Brain Sciences* 13, 5–13
96. Toney, A. and Caliskan, A. (2021) ValNorm Quantifies Semantics to Reveal Consistent Valence Biases Across Languages and Over Centuries. in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7203–7218

97. Sourati, Z. *et al.* (2026) The homogenizing effect of large language models on human expression and thought. *Trends in Cognitive Sciences* 0
98. Charlesworth, T.E.S. *et al.* (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
99. Cheryan, S. and Markus, H.R. (2020) Masculine defaults: Identifying and mitigating hidden cultural biases. *Psychological Review* 127, 1022–1052
100. Brady, W.J. *et al.* (2023) Algorithm-mediated social learning in online social networks. *Trends in Cognitive Sciences* 27, 947–960
101. Bailey, A.H. *et al.* (2025) Intersectional Male-Centric and White-Centric Biases in Collective Concepts. *Pers Soc Psychol Bull* 51, 2085–2102
102. Lewis, M. and Lupyan, G. (2020) Gender stereotypes are reflected in the distributional structure of 25 languages. *Nat Hum Behav* 4, 1021–1028
103. Charlesworth, T.E.S. *et al.* (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
104. Card, D. *et al.* (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
105. Kozlowski, A.C. *et al.* (2019) The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *Am Sociol Rev* 84, 905–949
106. 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
107. Pronin, E. *et al.* (2002) The Bias Blind Spot: Perceptions of Bias in Self Versus Others. *Pers Soc Psychol Bull* 28, 369–381
108. Pronin, E. and Kugler, M.B. (2007) Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *Journal of Experimental Social Psychology* 43, 565–578
109. Danbold, F. and Unzueta, M.M. (2020) Drawing the diversity line: Numerical thresholds of diversity vary by group status. *Journal of Personality and Social Psychology* 118, 283–306
110. Unzueta, M.M. *et al.* (2012) Diversity Is What You Want It to Be: How Social-Dominance Motives Affect Construals of Diversity. *Psychol Sci* 23, 303–309
111. Kraus, M.W. *et al.* (2017) Americans misperceive racial economic equality. *Proceedings of the National Academy of Sciences* 114, 10324–10331
112. 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*, pp. 77–91
113. Karkkainen, K. and Joo, J. (2021) FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. presented at the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1548–1558
114. Morales, A. *et al.* (2021) SensitiveNets: Learning Agnostic Representations with Application to Face Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 2158–2164

115. Brady, W.J. *et al.* (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
116. Brady, W.J. *et al.* (2023) Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility. *Nat Hum Behav* 7, 917–927
117. Kordzadeh, N. and Ghasemaghaei, M. (2022) Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems* 31, 388–409
118. 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
119. Fogliato, R. *et al.* (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*, pp. 100–111
120. Charlesworth, T.E.S. *et al.* (2022) Historical representations of social groups across 200 years of word embeddings from Google Books. *Proceedings of the National Academy of Sciences* 119, e2121798119
121. Baltrušaitis, T. *et al.* (2019) Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 423–443
122. Drissi, M. (2024) More is Less? A Simulation-Based Approach to Dynamic Interactions between Biases in Multimodal Models. Arxiv.
123. Ghate, K. *et al.* (2025) Biases Propagate in Encoder-based Vision-Language Models: A Systematic Analysis From Intrinsic Measures to Zero-shot Retrieval Outcomes. in *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18562–18580
124. Guilbeault, D. *et al.* (2024) Online images amplify gender bias. *Nature* 626, 1049–1055
125. Glikson, E. and Woolley, A.W. (2020) Human Trust in Artificial Intelligence: Review of Empirical Research. *ANNALS* 14, 627–660
126. 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
127. 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
128. 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
129. Stephany, F. and Duszynski, J. (2026) Women Worry, Men Adopt: How Gendered Perceptions Shape the Use of Generative AI. ArXiv.
130. Choung, H. *et al.* (2023) Trust in AI and Its Role in the Acceptance of AI Technologies. *International Journal of Human–Computer Interaction* 39, 1727–1739
131. Schoeffer, J. *et al.* (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*, pp. 1616–1628

132. Bigman, Y.E. and Gray, K. (2018) People are averse to machines making moral decisions. *Cognition* 181, 21–34
133. 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
134. Skitka, L.J. *et al.* (1999) Does automation bias decision-making? *International Journal of Human-Computer Studies* 51, 991–1006
135. Gerlich, M. (2025) AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies* 15, 6
136. Ma, X. (2022) Internet use and gender wage gap: evidence from China. *J Labour Market Res* 56, 15
137. Verma, A. *et al.* (2022) An investigation of skill requirements in artificial intelligence and machine learning job advertisements. *Industry and Higher Education* 36, 63–73
138. Kunda, Z. (1990) The case for motivated reasoning. *Psychological Bulletin* 108, 480–498
139. Sidanius, J. and Pratto, F. (1999) *Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression*, Cambridge University Press
140. Araujo, T. *et al.* (2020) In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Soc* 35, 611–623
141. Bertrand, A. *et al.* (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*, pp. 78–91
142. 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
143. Logg, J.M. *et al.* (2019) Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151, 90–103
144. Nickerson, R.S. (1998) Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology* 2, 175–220
145. Xuan, H. *et al.* (2026) Confirmation bias in the age of AI: Examining the role of information source and conclusive recommendations. *Computers in Human Behavior Reports* 22, 101047
146. Rathje, S. *et al.* (2026) The Impact of Sycophantic AI on Attitudes and Decisions. PsyArXiv.
147. Dietvorst, B.J. *et al.* (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 114–126
148. Dietvorst, B.J. *et al.* (2018) Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science* 64, 1155–1170
149. Tversky, A. and Kahneman, D. (1974) Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 1124–1131
150. Rastogi, C. *et al.* (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
151. Abdollahpouri, H. *et al.* (2020) The Connection Between Popularity Bias, Calibration, and Fairness in Recommendation. in *Proceedings of the 14th ACM Conference on Recommender Systems*, pp. 726–731

152. Feng, Y. and Chen, H. (2025) Algorithm-shaped norms: how YouTube's comment ranking algorithm influences consumer reactions to social issue campaigns. *Qualitative Market Research: An International Journal* 28, 778–801
153. Deutsch, M. and Gerard, H.B. (1955) A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology* 51, 629–636
154. Cialdini, R.B. *et al.* (1991) A Focus Theory of Normative Conduct: A Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior. In *Advances in Experimental Social Psychology* 24 (Zanna, M. P., ed), pp. 201–234, Academic Press
155. Duckitt, J. and Sibley, C.G. (2010) Personality, Ideology, Prejudice, and Politics: A Dual-Process Motivational Model. *Journal of Personality* 78, 1861–1894
156. Crandall, C.S. *et al.* (2018) Changing Norms Following the 2016 U.S. Presidential Election: The Trump Effect on Prejudice. *Social Psychological and Personality Science* 9, 186–192
157. Moberg, S.P. *et al.* (2019) Racial Attitudes in America. *Public Opin Q* 83, 450–471
158. Ruisch, B.C. and Ferguson, M.J. (2022) Changes in Americans' prejudices during the presidency of Donald Trump. *Nat Hum Behav* 6, 656–665
159. Kim, S. (2025) Perceptions of discriminatory decisions of artificial intelligence: Unpacking the role of individual characteristics. *International Journal of Human-Computer Studies* 194, 103387
160. Ghasemaghaei, M. and Kordzadeh, N. (2024) Understanding how algorithmic injustice leads to making discriminatory decisions: An obedience to authority perspective. *Information & Management* 61, 103921
161. Glickman, M. and Sharot, T. (2025) How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nat Hum Behav* 9, 345–359
162. Bender, E.M. *et al.* (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623
163. Edelman, B.L. *et al.* (2022) Inductive Biases and Variable Creation in Self-Attention Mechanisms. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 5793-5831.
164. Rahwan, I. (2018) Society-in-the-loop: programming the algorithmic social contract. *Ethics Inf Technol* 20, 5–14
165. Gebru, T. *et al.* (2021) Datasheets for datasets. *Commun. ACM* 64, 86–92
166. Kirk, H.R. *et al.* (2024) The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. *Advances in Neural Information Processing Systems* 37, 105236–105344
167. Fang, S. *et al.* (2025) SignLLM: Sign Language Production Large Language Models. in *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 6681–6693
168. Üstün, A. *et al.* (2024) Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model. in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15894–15939

169. Devine, P. *et al.* (2025) Prejudice Reduction. In *Handbook of Social Psychology 6th Edition* ((6th edn)) (Gilbert, D. et al., eds), Situational Press
170. Paluck, E.L. *et al.* (2021) Prejudice Reduction: Progress and Challenges. *Annual Review of Psychology* 72, 533–560
171. Amodio, D.M. *et al.* (2004) Neural Signals for the Detection of Unintentional Race Bias. *Psychol Sci* 15, 88–93
172. Devine, P.G. (1989) Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology* 56, 5–18
173. Devine, P.G. *et al.* (2012) Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology* 48, 1267–1278
174. Hildebrand, L.K. *et al.* (2024) Social Cognition and the Reduction of Intergroup Bias. In *The Oxford Handbook of Social Cognition, Second Edition* (Carlston, D. E. et al., eds), pp. 0, Oxford University Press
175. Amodio, D.M. and Swencionis, J.K. (2018) Proactive control of implicit bias: A theoretical model and implications for behavior change. *Journal of Personality and Social Psychology* 115, 255
176. Monteith, M.J. *et al.* (2002) Putting the brakes on prejudice: On the development and operation of cues for control. *Journal of Personality and Social Psychology* 83, 1029–1050
177. Sheeran, P. *et al.* (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
178. 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
179. Chater, N. and Loewenstein, G. (2023) The i-frame and the s-frame: How focusing on individual-level solutions has led behavioral public policy astray. *Behavioral and Brain Sciences* 46, 1–26
180. Lewis, N.A. (2023) Cultivating Equal Minds: Laws and Policies as (De)biasing Social Interventions. *Annual Review of Law and Social Science* 19, 37–52
181. Zhang, Y. *et al.* (2021) Causal Intervention for Leveraging Popularity Bias in Recommendation. in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 11–20
182. Sloane, M. and Wüllhorst, E. (2025) A systematic review of regulatory strategies and transparency mandates in AI regulation in Europe, the United States, and Canada. *Data & Policy* 7, e11
183. Kleinberg, J. *et al.* (2020) Algorithms as discrimination detectors. *Proceedings of the National Academy of Sciences* 117, 30096–30100
184. Matz, S.C. *et al.* (2024) Personality Science in the Digital Age: The Promises and Challenges of Psychological Targeting for Personalized Behavior-Change Interventions at Scale. *Perspect Psychol Sci* 19, 1031–1056
185. U.S. Equal Employment Opportunity Commission (2024) High Tech, Low Inclusion Diversity in the High Tech Workforce and Sector 2014-2022

186. 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
187. Hebl, M. *et al.* (2020) Modern Discrimination in Organizations. *Annual Review of Organizational Psychology and Organizational Behavior* 7, 257–282
188. Plenke, M. (2015) The Reason This "Racist Soap Dispenser" Doesn't Work on Black Skin. *Mic.* https://www.mic.com/articles/124899/the-reason-this-racist-soap-dispenser-doesn-t-work-on-black-skin
189. Martins, L.L. and Sohn, W. (2022) How Does Diversity Affect Team Cognitive Processes? Understanding the Cognitive Pathways Underlying the Diversity Dividend in Teams. *ANNALS* 16, 134–178
190. Gray, M.L. and Suri, S. (2019) *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass - Mary L. Gray, Siddharth Suri - Google Books*, Harper Business.
191. Raji, I.D. *et al.* (2020) Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33–44
192. Whittaker, M. *et al.* (2018) AI Now Report 2018AI Now Institute
193. Greenwald, A.G. and Banaji, M.R. (1995) Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review* 102, 4–27
194. Greenwald, A.G. and Banaji, M.R. (2017) The implicit revolution: Reconceiving the relation between conscious and unconscious. *American Psychologist* 72, 861–871
195. Amodio, D.M. (2025) A learning and memory account of impression formation and updating. *Nat Rev Psychol* 4, 417–432
196. 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