PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES OPEN ACCESS

People see more of their biases in algorithms

Begum Celiktutan[a], Romain Cadario[a], and Carey K. Morewedge[b,1]

Edited by Elke Weber, Princeton University, Princeton, NJ; received October 10, 2023; accepted March 19, 2024

Algorithmic bias occurs when algorithms incorporate biases in the human decisions on which they are trained. We find that people see more of their biases (e.g., age, gender, race) in the decisions of algorithms than in their own decisions. Research participants saw more bias in the decisions of algorithms trained on their decisions than in their own decisions, even when those decisions were the same and participants were incentivized to reveal their true beliefs. By contrast, participants saw as much bias in the decisions of algorithms trained on their decisions as in the decisions of other participants and algorithms trained on the decisions of other participants. Cognitive psychological processes and motivated reasoning help explain why people see more of their biases in algorithms. Research participants most susceptible to bias blind spot were most likely to see more bias in algorithms than self. Participants were also more likely to perceive algorithms than themselves to have been influenced by irrelevant biasing attributes (e.g., race) but not by relevant attributes (e.g., user reviews). Because participants saw more of their biases in algorithms than themselves, they were more likely to make debiasing corrections to decisions attributed to an algorithm than to themselves. Our findings show that bias is more readily perceived in algorithms than in self and suggest how to use algorithms to reveal and correct biased human decisions.

algorithm | algorithmic bias | bias blind spot | debiasing

Significance

Algorithms incorporate biases in the human decisions that comprise their training data, which can amplify and codify discrimination. We examine whether algorithmic biases can be used to reveal and help correct undetected biases of the human decision-makers on which algorithms are trained. We show that people see more of their biases in the decisions of algorithms than in their own decisions. Because algorithms reveal more of their biases, people are also more likely to correct their biases when decisions are attributed to an algorithm than to themselves. Recognizing bias is a crucial first step for people and organizations motivated to reduce their biases. Our findings illustrate how to use algorithms as mirrors to reveal and debias human decision-making.

Algorithms learn and incorporate biases in the human decisions on which they are trained (1–5). Algorithmic bias amplifies and codifies discrimination due to the scale with which algorithms are used in applications from deciding who is hired (1, 6) to who receives healthcare or bail (2, 7). Algorithmic biases also make human biases transparent that had been opaque when human decisions were unspecified or unaggregated (8, 9). When Amazon trained an algorithm on its past human hiring decisions, for example, the hiring algorithm revealed a gender bias that had previously escaped notice (10). We examine whether algorithmic biases can be used to help human decision makers recognize and correct for their biases.

People have access to the output of their intuitive decisions but lack access to the associative processes by which those decisions were made (11, 12). Because people assess bias in their decision-making by introspectively examining their decision-making processes, bias in the self often goes unrecognized. By contrast, people more readily detect biases in the decisions of others because others are judged by their decisions rather than their decision-making processes. The phenomenon that people more readily perceive bias in the decisions of others than in their own decisions is the bias blind spot (13–16).

We theorize that similar psychological processes lead people to perceive more of their biases in the decisions of algorithms than in their decisions. We propose that people are more able and motivated (17) to see their biases in algorithms because, like other people, the decision-making process of algorithms is opaque (18–20) and people perceive decisions made by algorithms like decisions made by other people (21, 22). People should thus use the same criteria to evaluate bias in algorithms as they use to evaluate bias in other people and be less threatened by and dismissive of bias in the decisions of algorithms than self, even when algorithms are trained on their decisions (23, 24). In nine preregistered experiments (N = 6,175), we find evidence that people perceive more of their biases (e.g., irrelevant effects of age, attractiveness, gender, and race on interpersonal judgments) in the decisions of algorithms than in their decisions. The revelatory effect of algorithms holds when research participants are given incentives that encourage them to reveal their true beliefs and discourage strategic self-presentation. We find that people are as likely to see biases in algorithms trained on their decisions as biases in the decisions of other people and algorithms trained on the decisions of other people. Furthermore, we show this revelatory effect of algorithms is driven by cognitive and motivated processes. It is larger for people more prone to the bias blind spot and larger when people are motivated to appear unprejudiced. Finally, we find that because people see more of their biases in algorithms, people are more likely to correct their biased decisions when those decisions are attributed

Author affiliations: [a]Rotterdam School of Management, Department of Marketing Management, Erasmus University, Rotterdam 3062 PA, The Netherlands; and [b]Questrom School of Business, Department of Marketing, Boston University, Boston, MA 02215

Author contributions: B.C., R.C., and C.K.M. designed research; B.C. and R.C. performed research; B.C. and R.C. analyzed data; and B.C., R.C., and C.K.M. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: morewedg@bu.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2317602121/-/DCSupplemental.

Published April 10, 2024.

to an algorithm than to themselves. Our findings show how to use algorithms to reveal and correct bias in human decision-making and provide evidence that the psychological processes used to perceive algorithms are scaffolded on the processes used to perceive other people.

Paradigm

We used a similar paradigm to test our hypotheses in all experiments (see *SI Appendix* for additional details). All participants rated a set of targets (i.e., Airbnb listings or rideshare drivers) that varied randomly on relevant attributes (e.g., ratings, number of reviews) and varied systematically on a potentially biasing irrelevant attribute (i.e., age, attractiveness, gender, race). In the first part of the experiment, participants sequentially rated each target on analog sliders (e.g., likelihood of renting, perceived driving ability) in one of two phases (A and B), with targets presented randomly without replacement.

In all experiments, we included two experimental conditions in which we showed participants a summary of their target ratings from phase B. In the "self" condition, we truthfully attributed those target ratings to the participant (e.g., "your ratings"). In the "self-trained algorithm" condition, we deceptively attributed those target ratings to an algorithm trained on other target ratings made by the participant (e.g., "predicted by an algorithm trained on your data from phase A"). In experiments 1 and 2, we added two "real self-trained algorithm" conditions in which we truthfully showed participants predicted phase B target ratings from a real algorithm trained on their phase A target ratings. In experiments 3 and 4, we added two conditions in which we presented participants with a summary of their target ratings from phase B, but we attributed their ratings to other participants in the experiment ("others" condition) or to an algorithm trained on the phase A target ratings of other participants in the experiment ("other-trained algorithm" condition).

All participants were then told about a research finding that explained how the irrelevant attribute might bias target ratings (e.g., age, attractiveness, gender, race) and participants reported the extent to which they perceived that "*you* [*the algorithm/other participants*]" showed the biasing tendency on a seven-point Likert scale with endpoints 1 (not at all) and 7 (very much). This absolute judgment of perceived bias was adapted from bias blind spot research (13, 15), but avoids a potential confound in comparative judgments of perceived bias between self and other (25). Perceived bias was positively correlated with the actual bias exhibited by individual participants in all nine experiments [r_{range} = 0.17 to 0.38; $r_{average}$ = 0.28 (95% CI = 0.24, 0.31)]; these correlations are high relative to correlations reported in bias blind spot papers comparing perceived and actual bias (e.g., r_{range} = −0.25 to 0.14) (15). See *SI Appendix*, Fig. S17.

In all experiments, we predicted that participants would perceive the biasing influence of the irrelevant attribute to be greater in the "self-trained algorithm" than "self" condition. Replicating previous research on the bias blind spot, we also expected perceived bias to be greater in the "others" condition than "self" condition.

People See More of Their Biases in Algorithms

Experiments 1 and 2. In experiments 1 and 2, we tested whether people see more of their racial and age biases when those biases are reflected in the decisions of algorithms than in their own decisions. In a one-factor between-subjects design, we randomly assigned participants to one of four conditions: self, self-trained algorithm, first real self-trained algorithm, and second real self-trained algorithm. In experiment 1 (N = 801, Prolific Academic), participants evaluated Airbnb listings varying (randomly) on star ratings and (systematically) on whether the hosts had distinctively African American or White names (26). Participants in self, self-trained algorithm, and first real self-trained algorithm conditions evaluated the renting likelihood of 10 Airbnb listings in phase A and 6 Airbnb listings in phase B. To eliminate the possible influence of suspicion, participants in the second real self-trained algorithm condition only evaluated the 10 Airbnb listings in phase A. Participants in the self and self-trained algorithm conditions next saw a summary judgment of their ratings in phase B that was attributed to self or algorithm, respectively. In the real self-trained algorithm conditions, the summary ratings for phase B that were shown to participants were predicted by a participant-level regression model trained on their phase A ratings using two regression coefficients. One coefficient was star rating (five-point scale). The other coefficient was race associated with hosts (African American or White). After viewing phase B summary ratings, all participants rated the perceived influence of racial bias on those ratings. To validate the real self-trained algorithm, we estimated a mixed effect regression comparing its predicted phase B ratings with all ratings made by participants who completed phase B, which revealed a strong average correlation (β = 0.75, t = 44.90, P < 0.001). In experiment 2 (N = 800, Prolific Academic), participants evaluated Uber drivers varying (randomly) on star ratings and (systematically) on whether the driver was young or old. The design was identical to experiment 1. This time, participants evaluated driving skills of different drivers in phase A and phase B. In the real self-trained algorithm conditions, the summary ratings for phase B were predicted by a participant-level regression model trained on participants' phase A ratings using star rating and age of the drivers (young or old) as coefficients. After viewing phase B summary ratings, all participants rated the perceived influence of age bias on those ratings. To validate the real self-trained algorithm, we estimated a mixed effect regression comparing its predicted phase B ratings with all ratings made by participants who completed phase B, which revealed a strong average correlation (β = 0.92, t = 80.56, P < 0.001). In both experiments, we only used deception in the self-trained algorithm conditions.

In experiment 1, we regressed the perceived influence of racial bias on three dummies for the algorithm conditions, with the self condition as the reference category, while controlling for actual racial bias. As preregistered, participants perceived more racial bias (β = 0.89, t = 5.60, P < 0.001) when their ratings were attributed to an algorithm (self-trained algorithm: M = 3.20, SE = 0.12) than to themselves (self: M = 2.29, SE = 0.11; Fig. 1*A*). Participants also perceived more racial bias in the ratings of real algorithms trained on their ratings than in their ratings for both the first (M = 3.42, SE = 0.13, β = 1.10, t = 6.73, P < 0.001) and second real self-trained algorithm conditions (M = 3.47, SE = 0.14, β = 1.18, t = 7.14, P < 0.001). By contrast, there was no difference in the perceived racial bias in ratings made by participants that were attributed to an algorithm (self-trained algorithm) and ratings predicted by real self-trained algorithms (respectively, P = 0.21, P = 0.08), or between the real self-trained algorithm conditions (P = 0.62). Additional results and robustness checks are reported in *SI Appendix*, Tables S10–S13.

In experiment 2, we regressed the perceived influence of age bias on three dummies for the algorithm conditions, with the self condition as the reference category, while controlling for actual age bias. As preregistered, participants perceived more age bias (β = 1.05, t = 6.61, P < 0.001) when their ratings were attributed to an algorithm (self-trained algorithm: M = 3.88, SE = 0.11) than to themselves (self: M = 2.83, SE = 0.11; Fig. 1*B*). Participants also perceived more age bias in the ratings of real algorithms

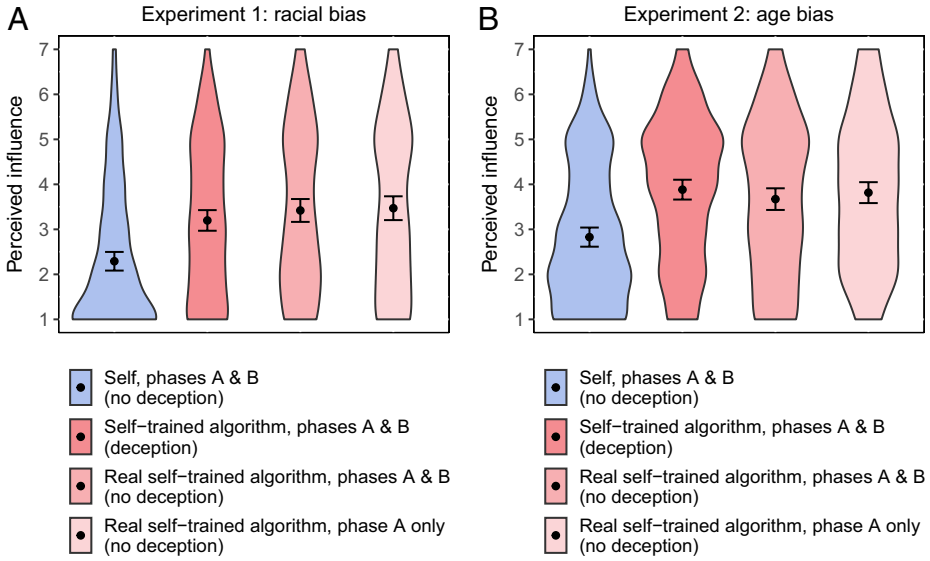

Fig. 1. People see more of their biases in algorithms (Experiments 1 and 2). The violin plots represent the shape of the distribution of perceived influence by experimental condition. The dot represents the mean, and the error bars represent the 95% CI. Experiment 1 is presented in panel *A* (N = 801), and Experiment 2 is presented in panel *B* (N = 800).

trained on their ratings than in their ratings for both the first (M = 3.67, SE = 0.12, β = 0.82, t = 5.15, P < 0.001) and second real self-trained algorithm conditions (M = 3.82, SE = 0.12, β = 0.92, t = 5.79, P < 0.001). By contrast, there was no difference in the perceived age bias in ratings made by participants that were attributed to an algorithm (self-trained algorithm) and ratings predicted by real self-trained algorithms (respectively, P = 0.16, P = 0.43), or between the real self-trained algorithm conditions (P = 0.53). Additional results and robustness checks are reported in *SI Appendix*, Tables S10–S13. Since perceived bias was similar between the self-trained algorithm and the real self-trained algorithm conditions in experiments 1 and 2, we used self-trained algorithm conditions in subsequent experiments so that the summary ratings of participants and algorithms were the same.

People See as Much of Their Biases in Algorithms as in Other People

Experiments 3 and 4. In experiments 3 and 4, we used a 2 (self, others) × 2 (participant, algorithm) between-subjects design. This design replicated our focal finding and tested if we would replicate the bias blind spot—people seeing more bias in the decisions of others than in their decisions. We recruited unique nationally representative online samples of US residents for each experiment from Prolific Academic. In experiment 3 (N = 797), each participant rated 18 Uber drivers (males and females) on driving skill in two phases of nine ratings (A and B). In experiment 4 (N = 775), each participant evaluated 18 Airbnb listings whose hosts had distinctively African American (9) or White (9) names, similar to experiment 1 (26). After rating all targets, participants saw a summary of target ratings in phase B. Then, participants rated the influence of gender or racial bias on those ratings (experiments 3 and 4, respectively).

In experiment 3, regressing the perceived influence of gender bias on self (0 for others, 1 for self), algorithm (0 for participant, 1 for algorithm), and their interaction while controlling for actual gender bias revealed the preregistered significant interaction (β = 0.85, t = 3.47, P < 0.001; Fig. 2*A*). Participants perceived more gender bias (β = 0.59, t = 3.40, P < 0.001) when driver ratings were attributed to an algorithm trained on their ratings (self-trained algorithm: M = 3.16, SE = 0.12) than to themselves (self: M = 2.62, SE = 0.12). By contrast, there was no difference in perceived bias (β = −0.26, t = −1.52, P = 0.13) whether driver ratings were attributed to an algorithm trained on other participants (other-trained algorithm: M = 3.43, SE = 0.13) or other participants (others: M = 3.64, SE = 0.13). Consistent with classic bias blind spot findings, participants perceived more gender bias when driver ratings were attributed to other participants than to themselves (others vs. self, β = −1.09, t = −6.34, P < 0.001). Additional results and robustness checks are reported in *SI Appendix*, Tables S10–S13.

In experiment 4, regressing the perceived influence of racial bias on the same predictors revealed the preregistered significant interaction (β = 0.82, t = 3.55, P < 0.001). Participants perceived more racial bias (β = 1.08, t = 6.62, P < 0.001; Fig. 2*B*) when listing ratings were attributed to an algorithm trained on their ratings (self-trained algorithm: M = 3.63, SE = 0.13) than to themselves (self: M = 2.47, SE = 0.12). By contrast, there was no difference in perceived bias (β = 0.26, t = 1.61, P = 0.109) whether listing ratings were attributed to an algorithm trained on other participants (other-trained algorithm: M = 3.89, SE = 0.12) or to other participants (others: M = 3.62, SE = 0.11). Consistent with classic bias blind spot findings, participants perceived more racial bias when listing ratings were attributed to other participants than to themselves (β = 1.12, t = 6.85, P < 0.001). Additional results and robustness checks are reported in *SI Appendix*, Tables S10–S13.

Why People See More of Their Biases in Algorithms

We tested whether cognitive and motivated drivers explain why people more readily perceive their biases in the decisions of algorithms than in their decisions in experiments 5 and 6.

Experiment 5. In experiment 5, we tested whether the revelatory effect of algorithms is moderated by individual differences in the bias blind spot, which is due to differences in the cognitive processes used to assess bias for self and others. People see less bias in their decisions than the decisions of others because they tend to introspectively look for biases in the process they used to make decisions (e.g., "I didn't think about gender when inviting speakers"). By contrast, because people lack introspective access to the decision processes of other people, they look for biases in the decisions made by other

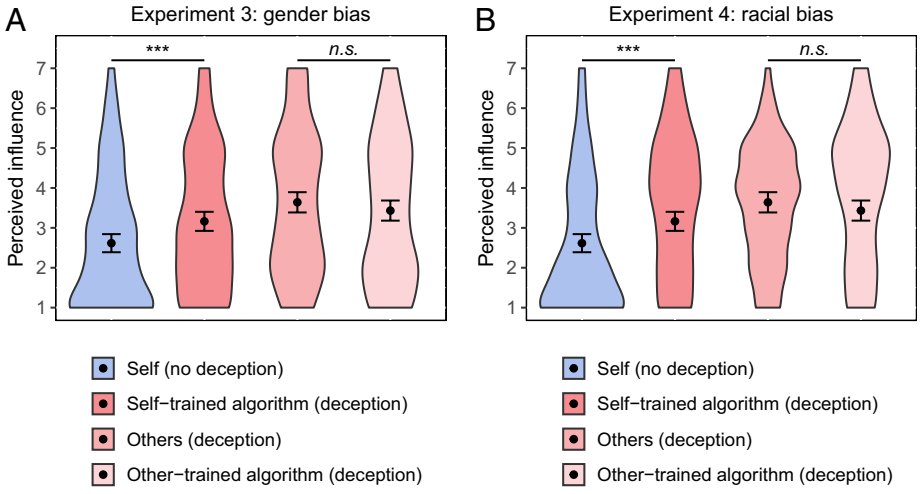

Fig. 2. People see as much bias in algorithms as in other people (Experiments 3 and 4). *** $P < 0.001$, "n.s." nonsignificant. The violin plots represent the shape of the distribution of perceived influence by experimental condition. The dot represents the mean, and the error bars represent the 95% CI. Experiment 3 is presented in panel *A* ($N = 797$), and Experiment 4 is presented in panel *B* ($N = 775$).

people (e.g., "All of their speakers are men"; ref. 15). People perceive decisions made by algorithms to be even more opaque (a "black box") than decisions made by other people (18, 19). Thus, differences in the tendency to exhibit the bias blind spot should moderate the perception of more bias in the decisions of algorithms than self. Participants ($N = 396$, Prolific Academic) were randomly assigned between-subjects to a self or self-trained algorithm condition and completed the same ratings and bias assessments as in experiment 3. All participants then completed a scale measuring susceptibility to the bias blind spot (15). We regressed perceived gender bias on condition (0 for self, 1 for self-trained algorithm), bias blind spot scale score, and their interaction while controlling for actual gender bias. The preregistered significant interaction ($\beta = 0.29$, $t = 2.12$, $P = 0.03$) revealed that susceptibility to bias blind spot increased the propensity to see more gender bias in ratings attributed to algorithm than to self. A floodlight analysis (27) revealed this difference was significant when bias blind spot scores were above 1.71 (*SI Appendix*, Fig. S18). Given proximity of this threshold to the mean ($M_{BBS} = 1.74$, $SE = 0.06$), we present dichotomized scores in Fig. 3*A*. See *SI Appendix*, Tables S10–S13 for additional results and robustness checks.

Experiment 6. We tested for the influence of motivated reasoning by examining whether algorithms selectively reveal the influence of biasing attributes in experiment 6. People are motivated to be unprejudiced for intrinsic and extrinsic reasons (28). If algorithms are perceived like other people, however, people should be less threatened by and dismissive of bias in decisions attributed to algorithms than self (23, 24). In a 2 (self, self-trained algorithm) × 2 (racial bias, star rating) between-subjects design, we manipulated whether participants ($N = 803$, Prolific Academic) reported the perceived influence of an attribute that would evoke a high or low motivation to respond without prejudice (28). Participants evaluated the likelihood of renting 18 Airbnb listings as in experiment 4. Half then reported the perceived influence of racial bias on target ratings made by self or self-trained algorithm (high motivation). Half were told that guests on the Airbnb platform are less likely to rent apartments from lower rated hosts than higher rated hosts and reported the perceived influence of star ratings on target ratings made by self or algorithm (low motivation). We regressed perceived influence on self-trained algorithm (0 for self, 1 for self-trained algorithm), racial bias (0 for star rating, 1 for racial bias), and

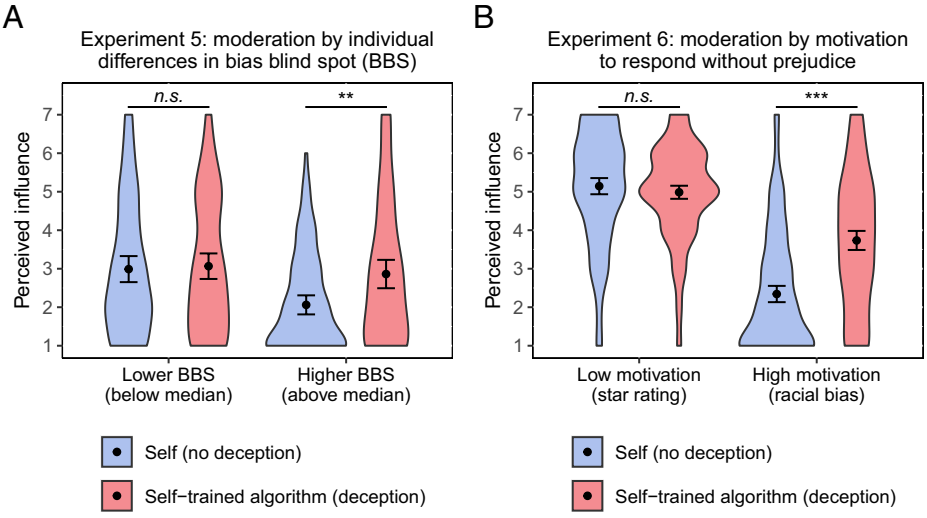

Fig. 3. Cognitive and motivated processes lead people to see more of their biases in algorithms (Experiments 5 and 6). ** $P < 0.01$, *** $P < 0.001$, "n.s." nonsignificant. The violin plots represent the shape of the distribution of perceived influence by experimental condition. The dot represents the mean, and the error bars represent the 95% CI. Experiment 5 is presented in panel *A* ($N = 396$), and Experiment 6 is presented in panel *B* ($N = 803$).

their interaction. The preregistered significant interaction revealed that algorithms selectively remove the bias blind spot (β = 1.55, t = 7.21, P < 0.001; Fig. 3*B*). Participants perceived more racial bias in ratings attributed to algorithms (β = 1.39, t = 9.17, P < 0.001; M = 3.74, SE = 0.13) than to themselves (M = 2.34, SE = 0.11). By contrast, participants perceived star ratings to have similarly influenced ratings attributed to algorithms (β = −0.16, t = −1.04, P = 0.297; M = 4.99, SE = 0.09) and to themselves (M = 5.14, SE = 0.11). See *SI Appendix*, Table S10–S13 for additional results and robustness checks.

People Are More Likely to Correct Their Biases in Algorithms

Experiment 7. We examined whether attributing decisions to algorithms makes people more willing to correct decisions in experiment 7. Participants (N = 400, Prolific Academic) were assigned to a 2 (self, self-trained algorithm; between-subjects) × 2 (precorrection, postcorrection; within-subjects) mixed design. All participants evaluated 18 Uber drivers in two phases that we varied systematically in facial attractiveness. Participants then reported the perceived biasing influence of attractiveness on driver ratings attributed to themselves or an algorithm trained on their ratings. Last, we allowed participants to correct driver ratings from phase B attributed to self or algorithm if they believed those ratings were biased. We computed a correction score, the average absolute difference in driver ratings before and after the opportunity for correction. As preregistered, participants corrected more when driver ratings were attributed to algorithms (β = 1.75, t = 3.28, P = 0.001; M = 3.77, SE = 0.46) than to themselves (M = 2.02, SE = 0.28). A similar difference in correction is observed when excluding outliers (β = 0.97, t = 4.03, P < 0.001; Fig. 4). Exploratory mediation analyses revealed that higher perceived influence of attractiveness bias predicted increased correction, which reduced actual attractiveness bias (*SI Appendix*, Figs. S20 and S21). See *SI Appendix*, Table S10–S13 for additional results and robustness checks.

Robustness Checks

Generalization across Heterogeneity in Actual Bias. We calculated a measure of actual bias in individual participant evaluations in all experiments (e.g., the average difference in evaluations of male and female Uber drivers). Descriptive analyses show that the distribution of actual bias exhibited by individual participants was heterogenous (*SI Appendix*, Fig. S16 and Table S9) and normally distributed around zero, except in experiment 7 and supplemental experiments where average bias was significantly greater than zero (all P's ≤ 0.002). About half of the participants in each sample exhibited bias in the direction as it was described to participants (e.g., favoring Whites to African Americans, favoring younger people, favoring males to females, favoring more attractive people). Importantly, variation in actual bias exhibited by individual participants did not moderate their propensity to perceive more of their bias in the decisions of algorithms than self in experiments 1 to 7 (*SI Appendix*, Table S12, Panel *A*). As an additional robustness check, among the participants who exhibited bias in the direction as it was described (e.g., favoring Whites to African Americans, males to females, more attractive people), the conditional effect of algorithm (vs. self) on perceived influence was significant in all experiments (*SI Appendix*, Table S12, Panel *B*).

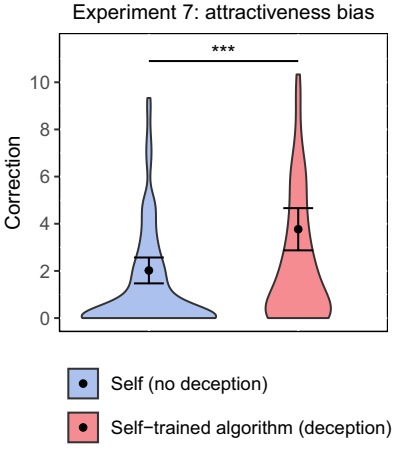

Fig. 4. People are more likely to correct their biases in algorithms (Experiment 7). ***P < 0.001. The violin plots represent the shape of the distribution of perceived influence by experimental condition. The dot represents the mean, and the error bars represent the 95% CI. The dependent variable is correction, excluding outliers above three times the interquartile range (N = 376). For additional figures without outlier exclusions, see *SI Appendix*, Fig. S19.

Generalization across Race and Gender. The race and gender of participants did not moderate the propensity to see more bias in algorithms than self in any experiment. In addition, self-identified race and gender did not consistently predict actual bias across experiments (*SI Appendix*, Table S13). Black participants were less likely to exhibit a bias favoring Whites to African Americans in experiment 4, for instance, but were no less likely in experiment 6.

Reflection of True Beliefs. In supplementary experiment A, half of participants were given a financial incentive (29) that encouraged them to reveal their true beliefs and discouraged strategic responding (30). In a 2 (self, real self-trained algorithm) × 2 (incentive, control) between-subjects design, participants (N = 800, Prolific Academic) evaluated the trustworthiness of attractive and unattractive Uber drivers. This experiment was modeled on experiments 1 and 2, using the second real self-trained algorithm condition in which participants only evaluated 10 drivers in phase A. To validate the real self-trained algorithm, we estimated a mixed effect regression comparing its predicted phase B ratings with all ratings made by participants who completed phase B, which revealed a strong average correlation (β = 0.86, t = 54.76, P < 0.001).

We regressed perceived influence on algorithm (−½ for self, ½ for real self-trained algorithm), incentive (−½ for control, ½ for incentive), and their interaction, while controlling for actual attractiveness bias (mean centered). There was a marginally significant main effect of incentive (β = 0.21, t = 1.79, P = 0.07) such that participants perceived more attractiveness bias in the incentive than control conditions. Exploratory analyses revealed that the increase in perceived attractiveness bias from control to incentive was marginally significant in the self condition (β = 0.31, t = 1.85, P = 0.07) but not significant in the real self-trained algorithm condition (β = 0.11, t = 0.68, P = 0.50; see *SI Appendix*, Fig. S22*A*). The interaction of algorithm and incentive was not significant (β = −0.20, t = −0.83, P = 0.41). Incentives may have increased perceived bias in self but did not moderate the propensity to see more bias in algorithms or the magnitude of the difference between self and algorithms. As predicted, participants perceived more attractiveness bias in ratings attributed to algorithms than to themselves (β = 0.91, t = 7.61, P < 0.001) both in control (M = 4.19, SE = 0.12 vs. M = 3.20, SE = 0.13; β = 1.01, t = 5.97, P < 0.001) and incentive conditions (M = 4.38, SE = 0.12 vs. M = 3.45, SE = 0.14; β = 0.81, t = 4.78, P < 0.001). See *SI Appendix*, Tables S10–S13 for additional results and robustness checks.

In supplementary experiment B, we tested whether incentivized ratings reflect true beliefs, not confusion about incentives, by using

a simpler incentive offered to all participants and including a comprehension check. The check revealed that 76% of the participants understood the incentive. We regressed the perceived influence of attractiveness bias on ratings made by self or a real self-trained algorithm (0 for self, 1 for real self-trained algorithm), while controlling for actual attractiveness bias (mean centered). Participants perceived more attractiveness bias in ratings made by the algorithm than self whether we included all participants in the analysis (M = 4.20 SE = 0.11 vs. M = 3.55, SE = 0.11; β = 0.63, t = 4.35, P < 0.001) or, as preregistered, only included participants who passed the comprehension check (M = 4.22, SE = 0.12 vs. M = 3.54, SE = 0.13; β = 0.63, t = 3.77, P < 0.001; see *SI Appendix*, Fig. S22*B*). In summary, participants appear to truly perceive more bias in decisions made by algorithms than themselves. See *SI Appendix*, Tables S10–S13 for additional results and robustness checks.

Discussion

Algorithms incorporate biases in the human decisions that comprise their training data, which can amplify and codify discrimination (1–5, 10). Our findings suggest auditing algorithms for bias can be beneficial not only for reducing algorithmic bias, but also for revealing biases in the human decisions on which they are trained. We find algorithms to be exempt from the bias blind spot that selectively inhibits people from recognizing and correcting their biased and prejudiced decision-making. In nine experiments, participants saw more of their biases in algorithms trained on their decisions than in their decisions (average d = 0.51; see *SI Appendix*, Fig. S15). For people and organizations motivated to reduce bias, recognizing bias is a crucial first step (31–34). Our findings present initial evidence that algorithms can serve as mirrors that reveal and debias human decision-making.

Materials and Methods

The present research involved no more than minimal risks, and all study participants were 18 y of age or older. All experiments were approved for use with human participants by the Institutional Review Board on the Charles River Campus at Boston University (protocol 3632E) or Institutional Review Board at Erasmus University (ETH2324-0356); informed consent was obtained for all participants. All manipulations and measures are reported. Experiments were conducted on the Qualtrics survey platform. Condition assignments were random in all our experiments, with randomization administered by Qualtrics. Preregistrations, surveys, raw data and reproducible R code are available on the Open Science Framework at https://osf.io/yvjt3/?view_only=6d6abf4759ea4bab9588d70c7b77c0d0.

Following a general rule of thumb, we sought to obtain a minimum of 200 participants per experimental condition. For each study, we requested the preregistered sample size on the online platform (i.e., N = 800 or N = 600 or N = 400). The final sample was determined by the actual number of participants who signed up for each online study, which was slightly higher or lower than the preregistered sample size. We preregistered N=800 for experiments 1, 2, 3, 4, and 6 and supplementary experiment A. The number of complete responses was respectively N = 801 (N = 801 total sample, that is 0% dropout), N = 800 (N = 800 total sample, that is 0% dropout), 797 (N = 813 total sample, that is 2.0% dropout), 775 (N = 775 total sample, that is 0% dropout), 803 (N = 803 total sample, that is 0% dropout), 800 (N = 804 total sample, that is 0.5% dropout). We preregistered N = 400 for experiments 5 and 7. The number of complete responses was respectively N = 396 (N = 396 total sample, that is 0% dropout), 400 (N = 400 total sample, that is 0% dropout). We preregistered N = 600 for supplementary experiment B. The number of complete responses was respectively N = 603 (N = 603 total sample, that is 0% dropout), The representative sampling for experiment 3 and 4 were performed by Prolific by matching the sample to the US population distribution by age, gender, and ethnicity. Balanced sample (i.e., even distributions of male and female participants) and a ≥98% approval rate were panel-related conditioning factors used in the other experiments.

Experiment 1. We recruited an online sample of 801 US residents (M_{age} = 35.9 y, 48% female) from Prolific Academic. Participants were randomly assigned to one of the four conditions: self, self-trained algorithm, first real self-trained algorithm, and second real self-trained algorithm. Participants imagined they were looking for a one-bedroom apartment to rent for a weekend. We presented information about apartments with a description of the apartment and the name of the host, using distinctively African American and distinctively White names (26). The information about each listing also included two diagnostic attributes (i.e., star rating and number of reviews), with randomly generated values for each participant. The diagnostic attributes are typically provided on platforms such as Airbnb. Under the apartment description, we kept the number of reviews constant for each apartment (i.e., 100 more), but for each apartment, we generated a set of 10 randomly generated numbers for star rating between 3.9 and 5. In other words, the star rating of an apartment varied randomly across participants. The full list of stimuli is available in *SI Appendix*.

Participants in self, self-trained algorithm, and first real self-trained algorithm conditions reported their likelihood of renting 10 apartments in phase A and 6 apartments in phase B. Participants in second real self-trained algorithm condition only reported their likelihood of renting 10 apartments in phase A. Participants rated their likelihood to rent each apartment with a 100-point analog slider scale with endpoints 0 (not at all likely) to 100 (very much likely). Importantly, we hid the slider values to participants. While they had a sense of low and high likelihoods, participants were unable to know their exact driving evaluation values. After evaluating the apartments, participants in the self condition moved directly to the dependent variable page, while participants from the algorithm conditions read additional information about an algorithm that an algorithm was said to use "your own" evaluation data from phase A to predict the evaluation from phase B. The algorithm information is presented in *SI Appendix*, Fig. S1.

On the dependent variable page, we presented participants with a summary table including six apartments from phase B, with African American and White host names grouped separately. Importantly, while participants in self and self-trained algorithm conditions were presented with *their own* ratings from phase B, participants in the real self-trained algorithm conditions viewed the summary evaluations for phase B predicted by an individual-level regression model on the 10 observations from phase A with two independent variables (a dummy for African American or White name, and a continuous predictor for the star rating) and with renting likelihood as dependent variable. In the first real self-trained algorithm condition, participants completed evaluations in phase B. In the second real self-trained algorithm condition, participants did not complete phase B. Below the summary evaluation table, we presented participants with a short statement about research on racial bias "*Research suggests that guests on the Airbnb platform are less likely to rent apartments from hosts with distinctly African American names than with distinctively White names.*" Finally, we measured our dependent variable of perceived influence with a single item: "to what extent do you believe that you (the algorithm) showed this tendency" on a seven-point Likert scale with 1 as not at all and 7 as very much, adapted from prior research on the bias blind spot (13, 15). Example evaluation pages are presented in *SI Appendix*, Figs. S8 and S9. Last, all participants reported age, gender, and ethnicity.

Experiment 2. We recruited an online sample of 800 US residents (M_{age} = 38.6 y, 49% female) from Prolific Academic. Participants were randomly assigned to one of the four conditions: self, self-trained algorithm, first real self-trained algorithm, and second real self-trained algorithm. Participants imagined they would use a ride-sharing service and evaluated different drivers. We presented information about drivers, which involved a photo from Chicago Face Database (35) and the two diagnostic attributes (i.e., star rating and number of reviews) identical to experiment 1. To create a young and an old version of each photo, we edited these photos with an AI tool (https://ailab.wondershare.com/tools/aging-filter.html). The design was similar to experiment 1, in which participants in self, self-trained algorithm, and first real self-trained algorithm conditions rated perceived driving skill of 10 drivers in phase A and six drivers in phase B. Participants in second real self-trained algorithm condition only rated perceived driving skill of 10 drivers in phase A. Participants rated drivers' driving skill with a 100-point analog slider scale with endpoints 0 (not at all skilled) to 100 (highly skilled). After evaluating the drivers, participants in the self condition moved directly to the dependent variable page, while participants from the algorithm conditions read additional

information about an algorithm that an algorithm was said to use "your own" evaluation data from phase A to predict the evaluation from phase B, similar to experiment 1. The algorithm information is presented in *SI Appendix*, Fig. S2.

On the dependent variable page, we presented participants with a summary table including six drivers from phase B, with young and old drivers grouped separately. Participants in self and self-trained algorithm conditions were presented with *their own* ratings from phase B, whereas participants in the real algorithm conditions viewed the summary evaluations for phase B predicted by an individual-level regression model on the 10 observations from phase A with two independent variables (a dummy for young or old driver, and a continuous predictor for the star rating) and with perceived driving skill as dependent variable. Below the summary evaluation table, we presented participants with a short statement about research on age bias "*Research on age biases suggests that people show a tendency to associate younger people with more driving skill than older people.*" We asked them to examine their driving skill evaluations from Phase B (the driving skill evaluations from Phase B predicted by an algorithm trained on their evaluations) for this age bias and measured perceived influence with the same scale as in experiment 1. Example evaluation pages are presented in *SI Appendix*, Figs. S10 and S11. Last, all participants reported age, gender, and ethnicity.

Experiment 3. We recruited a nationally representative online sample of 797 US residents (M_{age} = 45.78 y, 49% female) from Prolific Academic. Participants were randomly assigned to one of the four conditions in a 2 (self, others) × 2 (participant, algorithm) between-subjects design. Participants imagined they would use a ride-sharing service and evaluated different drivers. Then, we presented them information about 18 drivers (nine female and nine male) in two phases (i.e., phase A and phase B). The driver information presented involved a photo from Chicago Face Database (35) and four diagnostic attributes (i.e., number of trips, star rating, experience in platform and brand of the car). We chose the diagnostic attributes because they are typically provided on ridesharing platforms such as Uber. Under each photograph, we assigned each participant to a random selection of attribute values. In other words, the same driver had different attribute values across participants, similar to experiments 1 and 2. Every attribute included 10 different values. The 10 values were randomly generated numbers between 1,000 and 3,000 for the number of trips; randomly generated numbers between 4.00 and 5.00 for star rating; the experience in the platform ranged from 8 mo to 3.5 y; 10 car brands were selected from a list providing most popular cars commonly used by Uber drivers. The full list of stimuli is available in *SI Appendix*.

Participants rated every driver on perceived driving skill with a 100-point slider scale from 0 (not at all skilled) to 100 (very much skilled). After evaluating the 18 drivers in phases A and B, participants in the self and others conditions moved directly to the dependent variable page, while participants from the self-trained and other-trained algorithm conditions read additional information about an algorithm. In the self-trained algorithm condition, the algorithm was said to use "your own" evaluation data from phase A to predict the evaluation from phase B. In the other-trained algorithm condition, the algorithm was said to use evaluation data from phase A from "other participants of this study" to predict the evaluation from phase B. The algorithm information is presented in *SI Appendix*, Fig. S3.

On the dependent variable page, we presented participants with a summary table including the actual driving evaluation values for the nine drivers from phase B, ranked from highest to lowest. Participants in all conditions viewed *their own* driving evaluations from phase B, however, we provided different attributions across conditions. Participants in the self condition were informed that the table summarized their own evaluations. Participants in the others condition were informed that the table summarized the evaluations from other participants of the study. Participants in the self-trained algorithm conditions were informed that the table summarized the predicted evaluations by the algorithm trained on their own data. Participants in the other-trained algorithm conditions were informed that the table summarized the predicted evaluations by the algorithm trained on other participants' data. Below the summary evaluation table, we presented participants with a short statement about research on gender bias "*Research on gender biases suggests that people show a tendency to associate men with higher driving skills than women.*" Finally, we measured perceived influence with the same scale as in previous studies. Example evaluation pages are presented in *SI Appendix*, Figs. S12 and S13. Last, all participants reported age, gender, and ethnicity.

Experiment 4. We recruited a nationally representative online sample of 775 US residents (M_{age} = 45.19 y, 50% female) from Prolific Academic. Participants were randomly assigned to one of the four conditions in a 2 (self, others) × 2 (participant, algorithm) between-subjects design. Participants imagined they were looking for a one-bedroom apartment to rent for a weekend and evaluated the renting likelihood of apartments similar to experiments 1 and 2. We presented information about 18 apartments in two phases (i.e., phase A and phase B). In this experiment, we presented information about apartments that included a description of each listing and the name of its host, using distinctively African American and distinctively White names (26). The information about each listing also included four diagnostic attributes for each apartment (i.e., number of reviews, cleanliness star rating, communication star rating, and location star rating), with randomly generated values for each participant, similar to experiment 1. The list of attributes is presented in *SI Appendix*. The algorithm presentation and summary evaluation table were similar to experiment 3. Last, we informed participants about research on racial bias and measured its perceived influence with the same scale as in previous studies. Finally, participants reported age, gender, and ethnicity.

Experiment 5. We recruited an online sample of 396 US residents (M_{age} = 36.6 y, 47% female) from Prolific Academic. Participants were randomly assigned to either a self or a self-trained algorithm condition in a between-subjects design. Participants were assigned to the same tasks as presented in experiment 3. Next, participants responded to the 14-item Bias Blind Spot Scale (15), which measures individual differences in susceptibility to the bias blind spot. We calculated a bias blind spot score for each participant by subtracting the perceived susceptibility to bias of self from the perceived susceptibility to bias of average American for each bias and then averaging these differences (M_{BBS} = 1.73, SE = 0.06). Last, participants reported age, gender, and ethnicity.

Experiment 6. We recruited an online sample of 803 US residents (M_{age} = 36.9 y, 48% female) from Prolific Academic. Participants were randomly assigned to one of the four conditions in a 2 (self, self-trained algorithm) × 2 (racial bias, star rating) between-subjects design. The design was identical to experiment 4, in which participants evaluated the likelihood of renting of 18 apartments. In the racial bias condition, participants read *"Research suggests that guests on the Airbnb platform are less likely to rent apartments from hosts with distinctly African American names than with distinctively White names."* In the star rating condition, participants read "*Research suggests that guests on the Airbnb platform are less likely to rent apartments from lower rated hosts than higher rated hosts.*" Perceived influence was measured with the same scale used in previous experiments. Last, participants reported age, gender, and ethnicity.

Experiment 7. We recruited an online sample of 400 US residents (M_{age} = 37.2 y, 50% female) from Prolific Academic. Participants were assigned to a 2 (self, self-trained algorithm) × 2 (precorrection, postcorrection) mixed design; the first factor was manipulated between-subjects and the second factor was manipulated within-subjects. Participants were assigned to the same tasks as presented in experiment 3. Differently, they rated every driver on trustworthiness with a 100-point analog slider scale from 0 (not at all trustworthy) to 100 (very much trustworthy).

The algorithm presentation and summary evaluation table were similar to experiment 3. We informed participants about research on attractiveness biases "*Research on attractiveness biases suggests that people show a tendency to believe attractive people are more trustworthy than unattractive people.*" Then, we measured its perceived influence with the same scale as in other experiments. After, participants in the self condition were informed that they could change their evaluations from phase B, if they believed their evaluations were subject to an attractiveness bias. Participants in the self-trained algorithm condition were informed that they could change the algorithm's evaluations from phase B, if they believed its evaluations were subject to an attractiveness bias. After reading this information, participants rated each driver from phase B again, while we provided the original rating of the driver from phase B and the average rating for all drivers from phase B on the page. Also, the slider scale had a start position at their own evaluations from phase B. An example evaluation page is presented in *SI Appendix*, Fig. S14. Last, participants reported age, gender, and ethnicity.

Supplementary Experiment A. We recruited an online sample of 800 US residents from Prolific Academic (M_{age} = 37.72 y, 49% female). Participants were assigned to a 2 (self, real self-trained algorithm) × 2 (incentive, control) between-subjects design. Participants made the same ratings as in experiment 7, except that participants in the self conditions evaluated the trustworthiness of 10 drivers in phase A and 6 drivers in phase B. Participants in the real self-trained algorithm conditions only evaluated the trustworthiness of 10 drivers in phase A.

The algorithm presentation and summary evaluation table were similar to those of experiments 1 and 2; six attractive and unattractive drivers from phase B were grouped separately. Participants in self conditions were presented with their own ratings from phase B. Participants in the real self-trained algorithm conditions viewed the summary evaluations for phase B predicted by an individual-level regression model on the 10 observations from phase A with two independent variables (a dummy for attractive or unattractive driver, and a continuous predictor for the star rating) and with trustworthiness as dependent variable. We informed participants about research on attractiveness biases, "*Research on attractiveness biases suggests that people show a tendency to believe attractive people are more trustworthy than unattractive people.*" Then we measured its perceived influence with the same scale as in other experiments. Prior to reporting the perceived biasing influence of attractiveness on ratings, participants in incentive conditions read *"After this survey is concluded, we are going to run another study in which we show participants summary trust evaluations that vary in attractiveness bias. If, for trust evaluations with the same degree of bias as in the summary evaluations above, your perceived tendency rating matches their average rating, we will give you a $1 bonus."* Last, participants reported age, gender, and ethnicity.

Supplementary Experiment B. We requested an online sample of 600 US residents from Prolific Academic; 603 completed the experiment (M_{age} = 36.12 y, 47% female). The design of the experiment was identical to that of supplementary experiment A with three exceptions. First, all participants were incentivized; participants were randomly assigned to incentivized self or real self-trained algorithm conditions. Second, we used a simpler incentive, *"After you finish this study, we will show another participant the same trust evaluations. If your perceived tendency rating is the same number as their perceived tendency rating, we will give you a $1 bonus".* Third, we included a one-item comprehension check, *"Your perceived tendency rating was [PIPED TEXT]. What will the other participant's perceived tendency rating need to be for you to receive the $1 bonus?".* Participants responded on a seven-point scale with values from 1 to 7. Correct responses were coded as 1 and incorrect responses as 0.

Data, Materials, and Software Availability. Raw data and R code data have been deposited in Open Science Framework (https://osf.io/yvjt3/?view_only=6d6abf4759ea4bab9588d70c7b77c0d0) (36).

ACKNOWLEDGMENTS. C.K.M. gratefully acknowledges financial support from the Digital Business Institute in the Questrom School of Business at Boston University. B.C. and R.C. thank the Erasmus Research Institute of Management for their financial support for the data collection and the Research Software Engineering and Consulting team at Rotterdam School of Management for programming help.

1. A. Lambrecht, C. Tucker, Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Manage. Sci.* **65**, 2966–2981 (2019).
2. Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
3. V. Eubanks, *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor* (St. Martin's Press, 2018).
4. C. K. Morewedge *et al.*, Human bias in algorithm design. *Nat. Hum. Behav.* **7**, 1822–1824 (2023).
5. D. Danks, A. J. London, "Algorithmic bias in autonomous systems" in *Proceedings of 26th International Joint Conference on Artificial Intelligence*, C. Sierra, Ed. (International Joint Conferences on Artificial Intelligence, California, CA, 2017), pp. 4691–4697.
6. A. Bonezzi, M. Ostinelli, Can algorithms legitimize discrimination? *J. Exp. Psychol. Appl.* **27**, 447 (2021).
7. J. Angwin, J. Larson, S. Mattu, L. Kirchner, "Machine bias" in *Ethics of Data and Analytics* (Auerbach Publications, 2022), pp. 254–264.
8. J. Kleinberg, J. Ludwig, S. Mullainathan, C. R. Sunstein, Discrimination in the age of algorithms. *J. Leg. Anal.* **10**, 113–174 (2019).
9. J. Kleinberg, J. Ludwig, S. Mullainathan, C. R. Sunstein, Algorithms as discrimination detectors. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30096–30100 (2020).
10. J. M. Logg, "Using algorithms to understand the biases in your organization." *Harvard Business Review* (2019). https://store.hbr.org/product/using-algorithms-to-understand-the-biases-in-your-organization/H053H4. Accessed 10 October 2023.
11. C. K. Morewedge, D. Kahneman, Associative processes in intuitive judgment. *Trends. Cogn. Sci.* **14**, 435–440 (2010).
12. T. D. Wilson, N. Brekke, Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychol. Bull.* **116**, 117–142 (1994).
13. E. Pronin, D. Y. Lin, L. Ross, The bias blind spot: Perceptions of bias in self versus others. *Pers. Soc. Psychol. Bull.* **28**, 369–381 (2002).
14. E. Pronin, M. B. Kugler, Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *J. Exp. Soc. Psychol.* **43**, 565–578 (2007).
15. I. Scopelliti *et al.*, Bias blind spot: Structure, measurement, and consequences. *Manage. Sci.* **61**, 2468–2486 (2015).
16. E. Pronin, T. Gilovich, L. Ross, Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. *Psychol. Rev.* **111**, 781 (2004).
17. P. G. Devine, E. A. Plant, D. M. Amodio, E. Harmon-Jones, S. L. Vance, The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. *J. Pers. Soc. Psychol.* **82**, 835–848 (2002).
18. R. Cadario, C. Longoni, C. K. Morewedge, Understanding, explaining, and utilizing medical artificial intelligence. *Nat. Hum. Behav.* **5**, 1636–1642 (2021).
19. M. Yeomans, A. Shah, S. Mullainathan, J. Kleinberg, Making sense of recommendations. *J. Behav. Decis Mak.* **32**, 403–414 (2019).
20. A. Shariff, J.-F. Bonnefon, I. Rahwan, Psychological roadblocks to the adoption of self-driving vehicles. *Nat. Hum. Behav.* **1**, 694–696 (2017).
21. C. K. Morewedge, Preference for human, not algorithm aversion. *Trends. Cogn. Sci.* **26**, 824–826 (2022).
22. S. Agarwal, J. De Freitas, A. Ragnhildstveit, C. K. Morewedge, Acceptance of automated vehicles is lower for self than others. *J. Assoc. Consum. Res.*, in press.
23. W. K. Campbell, C. Sedikides, Self-threat magnifies the self-serving bias: A meta-analytic integration. *Rev. Gen. Psychol.* **3**, 23–43 (1999).
24. K. Kawakami, E. Dunn, F. Karmali, J. F. Dovidio, Mispredicting affective and behavioral responses to racism. *Science* **323**, 276–278 (2009).
25. D. R. Mandel, R. N. Collins, A. C. Walker, J. A. Fugelsang, E. F. Risko, Hypothesized drivers of the bias blind spot–Cognitive sophistication, introspection bias, and conversational processes. *Judgm. Decis. Mak.* **17**, 1392–1421 (2022).
26. M. Bertrand, S. Mullainathan, Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *Am. Econ. Rev.* **94**, 991–1013 (2004).
27. S. A. Spiller, G. J. Fitzsimons, J. G. Lynch Jr., G. H. McClelland, Spotlights, floodlights, and the magic number zero: Simple effects tests in moderated regression. *J. Mark. Res.* **50**, 277–288 (2013).
28. E. A. Plant, P. G. Devine, Internal and external motivation to respond without prejudice. *J. Pers. Soc. Psychol.* **75**, 811–832 (1998).
29. G. Charness, U. Gneezy, V. Rasocha, Experimental methods: Eliciting beliefs. *J. Econ. Behav. Organ.* **189**, 234–256 (2021).
30. E. P. Apfelbaum, S. R. Sommers, M. I. Norton, Seeing race and seeming racist? Evaluating strategic colorblindness in social interaction. *J. Pers. Soc. Psychol.* **95**, 918–932 (2008).
31. A.-L. Sellier, I. Scopelliti, C. K. Morewedge, Debiasing training improves decision making in the field. *Psychol. Sci.* **30**, 1371–1379 (2019).
32. K. L. Milkman, D. Chugh, M. H. Bazerman, How can decision making be improved? *Perspect. Psychol. Sci.* **4**, 379–383 (2009).
33. B. Fischhoff, "Debiasing" in *Judgment under Uncertainty* (Cambridge University Press, 1982), pp. 422–444.
34. S. P. Perry, M. C. Murphy, J. F. Dovidio, Modern prejudice: Subtle, but unconscious? The role of Bias Awareness in Whites' perceptions of personal and others' biases. *J. Exp. Soc. Psychol.* **61**, 64–78 (2015).
35. D. S. Ma, J. Correll, B. Wittenbrink, The Chicago face database: A free stimulus set of faces and norming data. *Behav. Res. Methods* **47**, 1122–1135 (2015).
36. B. Celiktutan, R. Cadario, C. K. Morewedge, Data from "People See More of Their Biases in Algorithms." OSF. https://osf.io/yvjt3/?view_only=6d6abf4759ea4bab9588d70c7b77c0d0. Deposited 3 March 2024.