Research Article

Psychological Science
1–12

DOI: 10.1177/0956797618805750
www.psychologicalscience.org/PS

People Make the Same Bayesian Judgment They Criticize in Others

Jack Cao[1], Max Kleiman-Weiner[1,2], and Mahzarin R. Banaji[1]
[1]Department of Psychology, Harvard University, and [2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

Abstract

When two individuals from different social groups exhibit identical behavior, egalitarian codes of conduct call for equal judgments of both individuals. However, this moral imperative is at odds with the statistical imperative to consider priors based on group membership. Insofar as these priors differ, Bayesian rationality calls for unequal judgments of both individuals. We show that participants criticized the morality and intellect of someone else who made a Bayesian judgment, shared less money with this person, and incurred financial costs to punish this person. However, participants made unequal judgments as a Bayesian statistician would, thereby rendering the same judgment that they found repugnant when offered by someone else. This inconsistency, which can be reconciled by differences in which base rate is attended to, suggests that participants use group membership in a way that reflects the savvy of a Bayesian and the disrepute of someone they consider to be a bigot.

Keywords

judgment, accuracy, fairness, social cognition, base rates, open data, open materials

Received 8/3/17; Revision accepted 8/3/18

O bir doktor. O bir hemsire.

The Turkish pronoun *o* is gender neutral, so these sentences translate to "One is a doctor. One is a nurse." However, Google's translation is "He is a doctor. She is a nurse." Although statistically accurate, because doctors are more likely to be men and nurses women, Google's translation raises the question of when it is appropriate to rely on group-based statistics. This question is relevant not just for developers of artificial intelligence (Caliskan, Bryson, & Narayanan, 2017) but also for anyone who makes social judgments while simultaneously attuned to group-based statistics (Garnham, Doehren, & Gygax, 2015) and norms concerning fairness (Shaw, 2016).

The appropriateness of relying on statistics is especially contentious when making judgments of individuals from different social groups who behave identically. Consider the following: A man performed surgery. A woman performed surgery. Who is more likely to be a doctor? The Bayesian answer is that the man is more likely to be a doctor because (a) doctors are more likely to be men and (b) not everyone who performs surgery is necessarily a doctor. Premise 1 is a well-known base rate, the neglect of which typifies a well-documented error (Kahneman & Tversky, 1973). Premise 2 may seem questionable if only doctors can perform surgery, but surgery includes procedures such as skin cancer excisions, which nurses can and do perform (Oliver, 2017).

The Bayesian answer is formalized in the following equation, which assesses how likely an individual (denoted as *target*) is to be a doctor given that he or she performed surgery:

$$\frac{P(\text{target} = \text{doctor} \mid \text{performed surgery})}{P(\text{target} \neq \text{doctor} \mid \text{performed surgery})} = \frac{P(\text{performed surgery} \mid \text{target} = \text{doctor})}{P(\text{performed surgery} \mid \text{target} \neq \text{doctor})} \times \frac{P(\text{target} = \text{doctor})}{P(\text{target} \neq \text{doctor})}$$

Corresponding Author:
Jack Cao, Harvard University, Department of Psychology, William James Hall, 33 Kirkland St., Room 1570, Cambridge, MA 02138
E-mail: jackcao@fas.harvard.edu

The prior, the bottom term, is greater when the target is a man rather than a woman. And because not everyone who performs surgery is necessarily a doctor, the likelihood, the middle term, will be large but less than infinity. If the likelihood does not depend on the target's gender, then the posterior, the top term, will be greater for a man than a woman. Thus, Bayesian rationality dictates that a man who performed surgery is more likely to be a doctor than a woman who performed surgery.

The representativeness heuristic (Kahneman & Tversky, 1972) also predicts that the man is more likely than the woman to be a doctor, conditional on both individuals performing surgery, because the man is more prototypical of the doctor profession. A Bayesian analysis makes two unique predictions that are tested. First, whereas the representativeness heuristic specifies only a direction, the Bayesian analysis, given values for the prior and likelihood, specifies a precise magnitude by which the man is more likely to be a doctor. Second, the Bayesian analysis is sensitive to variation in the likelihood, a term that the representativeness heuristic does not consider.

An alternative answer to the question of who is more likely to be a doctor is rooted in a moral imperative to make equal judgments of individuals who behaved identically (Dworkin, 2002; Rawls, 2001). Because both the man and the woman performed surgery, both should be viewed equally as doctors. This egalitarian ideal is embedded in more than 150 national constitutions, which state that men and women shall be treated equally (Constitute, 2016). Furthermore, promises of equal opportunity are common in the values statements of universities, corporations, and nonprofits. However, many researchers would agree that this aspiration is not always realized. Biernat and Kobrynowicz (1997) found that women are required to meet a higher threshold than men when demonstrating competence, a double standard associated with disparities in hiring, evaluation, and promotion (Eagly & Karau, 2002; Foschi, 1996; Rudman & Glick, 2001). This double standard has also been implicated in job applications (Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012), pay disparities (Auspurg, Hinz, & Sauer, 2017), and allegations of discrimination (*Price Waterhouse v. Hopkins*, 1989). Judging that the man is more likely than the woman to be a doctor when both individuals performed surgery can smack of yet another unfair double standard.

Under some circumstances, such as the Markov condition, in which base rates become irrelevant for maximizing statistical accuracy (Cao, Kleiman-Weiner, & Banaji, 2017), Bayesian rationality and morality are not in conflict. But in judging how likely a man, as opposed to a woman, is to be a doctor, given that they each performed surgery, there is a tension between the statistical and moral imperatives. The current work juxtaposes how people evaluate a third party that offers the Bayesian judgment against the judgment that people make themselves.

Driven by social desirability biases (Paulhus, 1991) or a genuine motivation to control prejudice (Plant & Devine, 1998), people may find someone who makes the Bayesian judgment to be morally flawed. Furthermore, the logic underlying the Bayesian judgment may be opaque, whereas egalitarian norms are so fundamental across many cultures (McAuliffe, Blake, Steinbeis, & Warneken, 2017) that even young children dislike unequal treatment of two individuals who did equal work (Shaw & Olson, 2014). These factors may lead participants to agree with the egalitarian judgment, a prediction supported by previous work that pits statistics against morality (Cao & Banaji, 2016). Unlike previous work, the current work primarily assessed how participants evaluate the morality and intellect of a third party that offers the Bayesian judgment, as well as how they treat this third party in economic games in which real money is at stake. Furthermore, the current work established boundary conditions by examining a wide range of professions.

Further extending previous work, the current research elicited probability judgments[1] and compared those judgments with what they should be, according to Bayes's theorem. If people condemn someone else for making the Bayesian judgment, then, to remain consistent (Abelson, 1968), people may not make the Bayesian judgment themselves. Given past research demonstrating deviations from Bayesian reasoning (Eddy, 1982) and statistical errors more generally (Tversky & Kahneman, 1974), it is also questionable whether people can even compute the Bayesian judgment. However, there is robust evidence suggesting that Bayesian prescriptions are apt descriptions of cognition in domains as wide ranging as object perception (Kersten, Mamassian, & Yuille, 2004), word learning (Xu & Tenenbaum, 2007), and everyday prediction (Griffiths & Tenenbaum, 2006). However, in none of these domains is morality a potential consideration. When statistics and morality are both at stake, what judgments do people make, and how do these judgments compare with how people evaluate someone else who makes the Bayesian judgment?

Study 1

Study 1 assessed how people evaluate a third party that offered the Bayesian judgment in the male-dominated profession of doctor.

Participants

One hundred ninety-nine participants (age: $M = 35.62$ years, $SD = 12.16$; 95 men, 104 women) were recruited from Amazon Mechanical Turk (MTurk) and compensated \$0.21 each. For all studies reported, as many participants were sampled as resources allowed, resulting in at least 100 participants per condition.

Procedure

Participants learned that a man had performed surgery and that a woman had performed surgery, after which participants indicated which of three statements they agreed with: (a) The man is less likely to be a doctor than the woman, (b) the man and woman are equally likely to be a doctor, or (c) the man is more likely to be a doctor than the woman.

Next, participants learned about person *X*, who stated the following after learning the same information as participants: "Even though the man and the woman both performed surgery, the man is more likely to be a doctor than the woman." Participants then completed four Likert-type scales that assessed how (a) fair, (b) just, (c) accurate, and (d) intelligent person *X*'s statement was. Each scale ranged from 1 (e.g., *extremely unfair*) to 7 (e.g., *extremely fair*).

Last, participants provided open-ended text responses of their impressions of person *X* and his statement. Throughout the procedure, the order in which the man and woman were compared was randomly assigned: For half the participants, stimuli stated that the man is more likely to be a doctor than the woman, whereas for the other half, stimuli stated that the woman is less likely to be a doctor than the man (for stimuli for all studies, see the Supplemental Material available online).

Results

A majority of participants (93%) agreed with the egalitarian judgment that the man and woman are equally likely to be a doctor, whereas 7% agreed with the Bayesian judgment that the man is more likely to be a doctor (see Fig. 1a). Critically, participants negatively evaluated person *X*, who was viewed as not only unfair, $M = 2.11$, $SE = 0.11$, and unjust, $M = 2.27$, $SE = 0.11$, but also inaccurate, $M = 2.31$, $SE = 0.12$, and unintelligent, $M = 2.41$, $SE = 0.11$, for making the Bayesian judgment, as indicated by means below the midpoint of 4 on the 7-point scales (see Fig. 1b), Cronbach's $\alpha = .92$, composite $M = 2.28$, $SE = 0.10$, one-sample $t(198) = -17.24$, $p < .0001$, Cohen's $d = 1.22$, 95% confidence interval (CI) = [0.99, 1.53]. Even though person *X*'s judgment was statistically accurate, this third party was negatively evaluated.

Study 2

Study 2 assessed how people evaluate a third party that offered the Bayesian judgment in a wide range of male-dominated professions aside from doctor.

Participants

Six hundred four participants (age: $M = 34.56$ years, $SD = 10.68$; 222 men, 382 women) were recruited from MTurk and compensated \$0.21 each. The procedure was identical to the procedure in Study 1 except for the professions in question and the behaviors exhibited by the man and woman.

Procedure

Each participant was randomly assigned to learn one of the following sets of information: (a) A man carved up a pig and a woman carved up a pig, (b) a man extinguished a fire and a woman extinguished a fire, or (c) a man poured concrete and a woman poured concrete. After learning this information, participants indicated whether they agreed that the man is less likely to be a butcher (or firefighter or construction worker), that the man and woman are equally likely to be a butcher (or firefighter or construction worker), or that the man is more likely to be a butcher (or firefighter or construction worker).

Participants then learned about person *X*, who, after learning the same information as participants, made the Bayesian judgment that the man is more likely to be a butcher (or firefighter or construction worker). Participants then completed four Likert-type scales that assessed how (a) fair, (b) just, (c) accurate, and (d) intelligent person *X*'s statement was. Each scale ranged from 1 (e.g., *extremely unfair*) to 7 (e.g., *extremely fair*). Last, participants provided open-ended text responses of their impressions of person *X* and this person's statement. Throughout the procedure, as in Study 1, the order in which the man and woman were compared was randomly assigned.

Results

Agreement with the Bayesian judgment that the man is more likely to be a butcher, firefighter, or construction worker was at least 33% (see Fig. 1a). Furthermore, evaluations of person *X* along the four dimensions of fair, just, accurate, and intelligent were neutral or slightly negative (see Fig. 1b). Given the high reliability

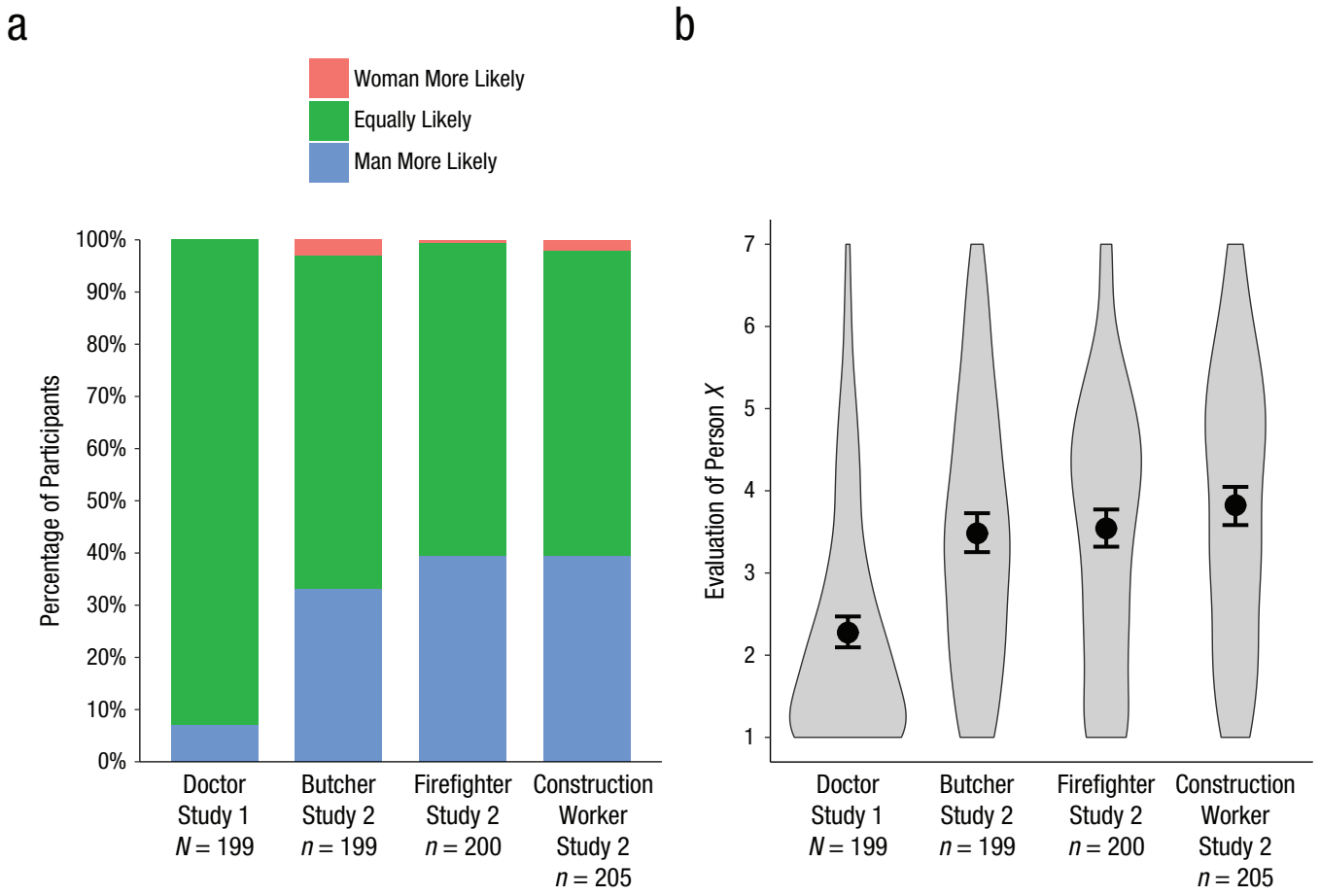

Fig. 1. Results from Studies 1 and 2: percentage of participants who agreed with each judgment (a) and evaluations of person *X* (b). Evaluations were computed by averaging the items "fair," "just," "accurate," and "intelligent," which were measured using Likert-type scales ranging from 1 to 7. Violin plots display the distributions, and dots indicates the means. Error bars show 95% confidence intervals.

of these four items within each condition, Cronbach's alphas were greater than .90, and all four items were averaged in each condition (for item means and standard errors, see Table S1 in the Supplemental Material). These average evaluations were just below the midpoint of 4, *M*s > 3.48, *SE*s < 0.12, and did not significantly differ from one another, Tukey adjusted *p*s > .08, *r*s < .08, indicating that a third party that offered the Bayesian judgment for these different professions was not condemned.

These results show that negative evaluations of a third party making a Bayesian judgment are circumscribed as opposed to universal. When the profession was doctor—but not butcher, firefighter, or construction worker—evaluations of person *X* were strongly negative. Base-rate differences may account for some of this variability: Whereas approximately two thirds of doctors in the United States are men, the proportion of construction workers who are men is greater. However, base rates alone cannot account for these results because the gender distribution among butchers is comparable with the gender distribution among doctors (Rocheleau, 2017). Other possibilities, therefore, include differences in status and the strength of norms regarding gender equality: There are efforts to increase the representation of women across various professions, but these efforts might be more pronounced among high-status science, technology, engineering, and math professions such as doctor. Nonetheless, person *X*'s Bayesian judgment was statistically accurate across all professions. Only when the profession was doctor did participants find extreme fault with person *X*'s morality and intellect for making the Bayesian judgment.

Study 3

Study 3 stringently tested whether evaluations of person *X* would remain negative when the profession is doctor. In Study 1, base rates and the possibility of nurses performing surgery were not salient. To increase the

salience of this information, we used a vignette in Study 3 in which (a) only one person, either the man or the woman but not both, could be the doctor, (b) whoever was not the doctor was a nurse, and (c) the framing was in terms of numerical percentages. Furthermore, Study 3 tested the behavioral implications of negatively evaluating person *X* by including an economic game in which participants were given the opportunity to share real money with this third party.

Participants

Four hundred twenty-five participants were recruited from MTurk. Each participant was compensated $0.21 and could have earned up to $0.30 more. Twenty-three participants were excluded for not completing the procedure. The final sample consisted of 402 participants (age: $M = 34.24$ years, $SD = 10.74$; 153 men, 248 women, 1 unspecified).

Procedure

Participants were instructed to imagine a man and a woman who work at the same hospital. One person is a doctor and the other person is a nurse, but who is the doctor and who is the nurse is unknown. Participants were instructed to assume that the man had performed surgery, in which case the probability that the man is the doctor is an unknown percentage. Participants were then instructed to assume that the woman had performed surgery, in which case the probability that the woman is the doctor is another unknown percentage (the order of these instructions was counterbalanced). Participants indicated whether they agreed that (a) the two percentages differ in that the man is less likely to be the doctor, (b) the two percentages are equivalent, or (c) the two percentages differ in that the man is more likely to be the doctor. As before, the order in which the man and woman were compared was counterbalanced.

Next, participants learned about person *X*, who, on the basis of random assignment, agreed with the Bayesian judgment that the two percentages differ in that the man is more likely to be the doctor or agreed with the egalitarian judgment that the two percentages are equivalent. Participants then completed four Likert-type scales that assessed how (a) fair, (b) just, (c) accurate, and (d) intelligent person *X*'s statement was. Each scale ranged from 1 (e.g., *extremely unfair*) to 7 (e.g., *extremely fair*). Last, participants provided open-ended text responses of their impressions of person *X*.

Finally, participants were endowed with $0.30 and could transfer any amount to person *X*. If negative evaluations of person *X* have behavioral implications, then participants would transfer less money to person *X* when this third party offers the Bayesian judgment relative to when this third party offers the egalitarian judgment. Participants kept the money that they chose not to transfer, and 2 randomly selected participants from a previous version of this study that did not include an economic game—one who agreed with the Bayesian judgment and another who agreed with the egalitarian judgment—received the transferred money.

Results

As observed before, a majority of participants (91%) agreed with the egalitarian judgment. Six percent agreed with the Bayesian judgment, and 3% agreed that the two percentages differ in that the woman is more likely to be the doctor. Further replicating previous negative evaluations of person *X*, results showed that this third party was again viewed as unfair, unjust, inaccurate, and unintelligent (for item means and standard errors, see Table S2 in the Supplemental Material) when the Bayesian judgment was offered, as indicated by means below the midpoint of 4 on the scales, Cronbach's $\alpha = .93$, composite $M = 2.49$, $SE = 0.10$, one-sample $t(201) = -14.78$, $p < .0001$, Cohen's $d = 1.04$, 95% CI = [0.83, 1.29]. This effect was reversed (see Fig. S1 in the Supplemental Material) when person *X* offered the egalitarian judgment; this version of person *X* was viewed as fair, just, accurate, and intelligent, as indicated by means above the midpoint of 4 on the scales, Cronbach's $\alpha = .91$, composite $M = 6.34$, $SE = 0.08$, one-sample $t(199) = 31.12$, Cohen's $d = 2.20$, 95% CI = [1.74, 2.88].

Critically, these evaluations have behavioral implications with real money. Participants transferred less money to person *X* when the Bayesian judgment was offered, $M = \$0.04$, $SE = \$0.004$, compared with when the egalitarian judgment was offered, $M = \$0.10$, $SE = \$0.01$, $b = -\$0.06$, $t(379.91) = 8.41$, $p < .0001$, Cohen's $d = 0.84$, 95% CI = [0.63, 1.04]. Also telling are the distributions of transfer amounts (see Fig. 2). When person *X* offered the egalitarian judgment, 54% of participants transferred at least half their endowment and 31% transferred nothing; when person *X* offered the Bayesian judgment, these percentages were reversed: 17% transferred at least half their endowment and 65% transferred nothing.

In the Supplemental Material, we report a study in which this behavioral result with real money was conceptually replicated; participants also willingly incurred a financial cost on themselves to punish someone who made the Bayesian judgment rather than the egalitarian judgment (see Table S3 and Figs. S2 and S3 in the Supplemental Material). These findings lend further credence to the behavioral implications of negative evaluations.

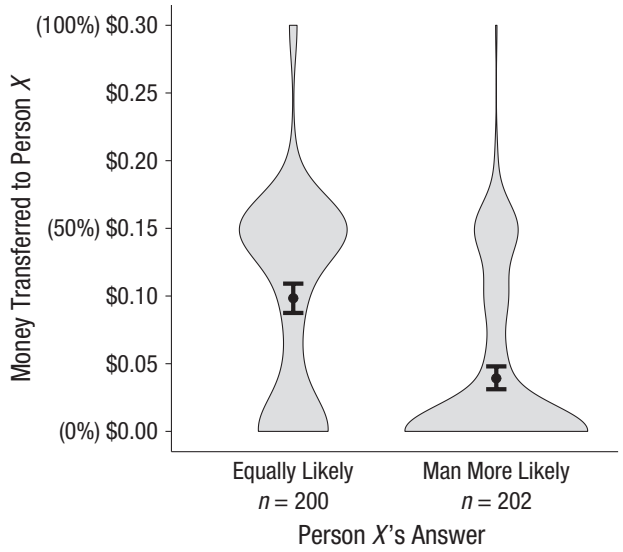

Fig. 2. Results from Study 3: average amounts transferred to person *X*. Violin plots display the distribution of amounts transferred in each condition, and dots indicates the means. Error bars show 95% confidence intervals.

Another study in the Supplemental Material shows that these effects are not simply due to the phrase "more likely" or "less likely," which may imply a large gap. The results were replicated when these phrases were precisely quantified as "8 percentage points more likely" or "8 percentage points less likely," a quantification that also increases the salience of base rates. A further study in the Supplemental Material conceptually replicated these results in a different profession, as participants also found fault with person *X* for making the Bayesian judgment that a man who communicated with air traffic control during a flight is more likely to be a pilot than a woman who communicated with air traffic control during a flight (see Table S4 and Fig. S4 in the Supplemental Material).

Study 4

Study 4 assessed whether probability judgments would be Bayesian or egalitarian. After learning that a man or a woman performed surgery, will participants judge that the man is more likely than the woman to be the doctor, or will participants judge that they are equally likely to be the doctor?

Participants

Eight hundred ninety-nine participants were recruited from MTurk and compensated $0.50 each. Some participants were randomly assigned to learn that a man had performed surgery, whereas others were randomly assigned to learn that a woman had performed surgery. This between-subjects design better approximates the conditions under which judgments are typically made.

Procedure

Three participants were excluded because they provided priors of either 0% or 100%, which cannot be updated according to Bayes's rule. Two participants indicated that they had looked up answers to some of the questions in the study, but these participants were retained in the analyses (conclusions do not change on the basis of whether these participants are included or excluded). Although it is possible that some participants looked up information but did not report doing so, this is not a problem for two reasons. First, there is considerable variability in participants' priors and likelihoods, which is consistent with participants drawing on their subjective beliefs as opposed to looking up information. Second, the critical questions involve Bayesian updating, so these questions are not easily answered through a search engine. The final sample consisted of 896 participants (age: $M = 34.33$ years, $SD = 10.97$; 528 men, 364 women, 4 unspecified). Study 4 proceeded in three parts, each corresponding to a component of Bayes's rule.

Part 1: priors. Participants were instructed to imagine a man and a woman who work at the same hospital. One person is a doctor and the other person is a nurse, but who is the doctor and who is the nurse is unknown. Participants estimated the percentage chance that each person is the doctor. Because there are two hypotheses—either the man or the woman is the doctor (and the other is the nurse)—both estimates had to sum to 1. Thus, each participant provided his or her subjective prior about each person's profession (e.g., the man has a 75% chance of being the doctor; the woman has a 25% chance of being the doctor).

Part 2: posteriors. After providing priors, each participant was randomly assigned to learn one of the following six pieces of data: (a) The man performed surgery on a patient, (b) the woman performed surgery on a patient, (c) the man gave a sponge bath to a patient, (d) the woman gave a sponge bath to a patient, (e), the man performed cardiopulmonary resuscitation (CPR) on a patient, and (f) the woman performed CPR on a patient. After learning this datum, participants again estimated the percentage chance that each person was the doctor. Thus, each participant provided his or her subjective posterior.

Performing surgery was chosen because it is highly diagnostic of the person being the doctor. Giving a

sponge bath was chosen because it is highly diagnostic of the person being the nurse. Performing CPR was chosen because it is relatively nondiagnostic of profession, as both doctors and nurses administer this procedure. For the primary analysis, only results from the surgery conditions (a and b) are presented. Data from the other four conditions are presented in Figures S5 and S6 in the Supplemental Material.

Part 3: likelihoods. Each participant estimated two likelihoods: the likelihood of observing the datum given the hypothesis that the target they learned about is the doctor and the likelihood of observing the datum given the hypothesis that the target they learned about is the nurse. For example, if a participant learned that the woman had performed surgery, that participant estimated the percentage of female doctors who perform surgery and the percentage of female nurses who perform surgery. If a participant learned that the man had performed surgery, that participant estimated the percentage of male doctors who perform surgery and the percentage of male nurses who perform surgery. Thus, each participant provided his or her subjective likelihood estimates, which were combined by forming a ratio. Each participant was randomly assigned to estimate the corresponding likelihoods either before or after providing subjective priors and posteriors.

Each participant's priors and likelihoods were entered into Bayes's rule to compute a *model posterior*, which represents what the participant's posterior should be from a statistical perspective. This model posterior was compared with the posterior that the participant actually reported.

Results

Participants' priors and likelihoods are discussed first before examining the correspondence between model and reported posteriors. When the target was a man, he was judged more likely to be the doctor a priori than when the target was a woman, man: $M = 68.7\%$, woman: $M = 29.6\%$, $b = 0.39$, $t(890) = 17.23$, $p < .0001$, $r = .50$, 95% CI = [.44, .52], as 81% of participants reported priors that favored the man over the woman to be the doctor.

As expected, likelihoods reflected the fact that not everyone who performs surgery is necessarily a doctor. Regardless of the gender of the target who performed surgery, the majority of participants indicated that some percentage of nurses perform surgery, resulting in likelihoods less than infinity. Furthermore, only a small difference in likelihoods was observed between the two conditions, man: $Mdn = 1.98$, woman: $Mdn = 2.65$; Wilcoxon $p = .19$, $r = .08$, which suggests that participants may have found the datum of performing surgery to be equally diagnostic of being a doctor, irrespective of the target's gender (see Fig. 3a). Many participants (< 41% in both conditions) found the datum of performing surgery to be entirely diagnostic, as shown by likelihoods equal to infinity. For these participants, their model posteriors are 100%, and their data are included in subsequent analyses of model and reported posteriors.

Because priors favored the man to be the doctor and because likelihoods were similar between the two conditions, model posteriors favored the man over the woman to be the doctor, even though both targets had performed surgery on a patient, man: $M = 87.7\%$, woman: $M = 72.2\%$, $b = 0.15$, $t(890) = 6.83$, $p < .0001$, $r = .22$, 95% CI = [.15, .27]. This disparity was also observed among participants' reported posteriors, man: $M = 86.4\%$, woman: $M = 78.0\%$, $b = 0.08$, $t(890) = 3.74$, $p = .0002$, $r = .12$, 95% CI = [.05, .18].

In fact, relatively small differences were observed between model and reported posteriors among participants who learned that the man had performed surgery, model posterior: $M = 87.7\%$, reported posterior: $M = 86.4\%$, $b = 0.01$, $t(1780) = 0.69$, $p = .49$, $r = .02$, 95% CI = [.001, .05], and among participants who learned that the woman had performed surgery, model posterior: $M = 72.2\%$, reported posterior: $M = 78.0\%$, $b = -0.06$, $t(1780) = -3.05$, $p = .002$, $r = .07$, 95% CI = [.006, .16]. Thus, the posteriors reported by participants were close to the posteriors they should have reported, according to Bayesian rationality (see Fig. 3b).

This close correspondence between model and reported posteriors suggests that participants integrated priors and likelihoods, as a Bayesian statistician would, and did not simply use the representativeness heuristic. Additional analyses in the Supplemental Material show (a) this close correspondence at the level of the individual participant, (b) the sensitivity of reported posteriors to likelihood ratios, and (c) that the critical comparisons hold when participants' probability judgments are logit transformed with a wide range of adjustment factors (see Figs. S7 and S8 in the Supplemental Material).

An additional study in the Supplemental Material further demonstrates that this effect generalizes to the profession of pilot, as participants judged that a man who communicated with air traffic control during a flight is more likely than a woman who communicated with air traffic control during a flight to be a pilot (see Figs. S9–S13 in the Supplemental Material). Together, these results indicate that participants' judgments reflect the statistical savvy of a Bayesian.

Study 5

Study 5 had a within-subjects design in which the same participants evaluated person *X* and made probability

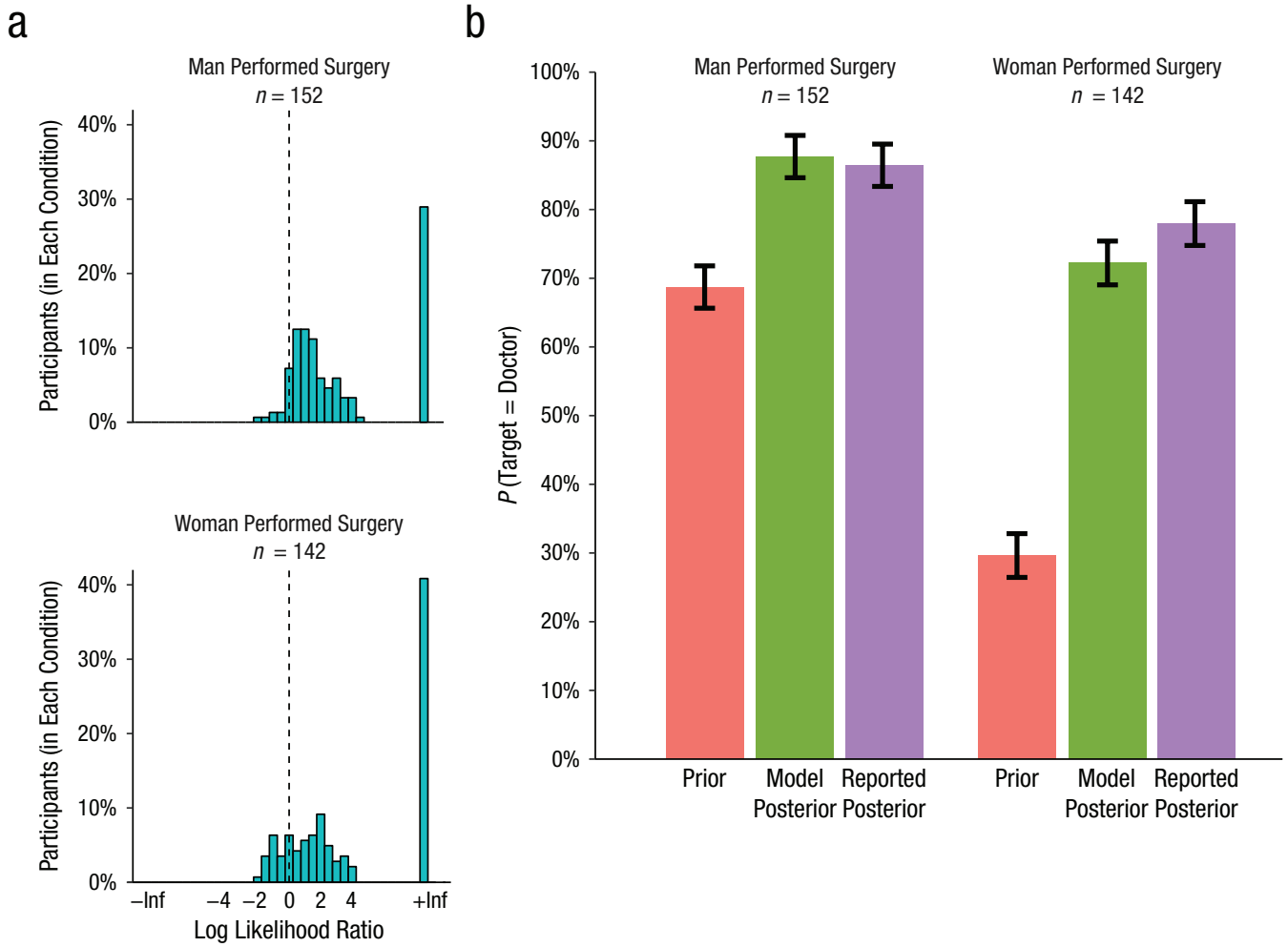

Fig. 3. Results from Study 4. The distributions (a) show likelihood ratios (log scaled) in each surgery condition. "Inf" refers to infinity, either negative (–) or positive (+). The graph (b) shows average judgments of participants in each surgery condition. "Prior" indicates judgments before participants learned that the target had performed surgery. "Model posterior" indicates judgments that participants should have made, from a statistical perspective, after learning that the target had performed surgery. "Reported posterior" indicates judgments that participants actually made after learning that the target had performed surgery. Error bars show 95% confidence intervals.

judgments. In this study, there would be demands to respond consistently, but failure to meet these demands would demonstrate how the same individual can make a Bayesian judgment but condemn someone else for making a Bayesian judgment.

Participants

Three hundred fifty-three participants were recruited from MTurk and compensated $0.71 each. Five participants were excluded because they provided priors that could not be updated according to Bayes's rule. Twenty-eight participants indicated that they had looked up answers to some of the questions in the study, but these participants were retained in the analyses (conclusions do not change on the basis of whether these participants are included or excluded; the higher number of participants who reported looking up answers is due to the inclusion of filler tasks consisting of trivia questions). Although it is possible that some participants looked up information but did not report doing so, this is not a problem for the same reasons discussed in Study 4. The final sample consisted of 348 participants (age: $M = 36.28$ years, $SD = 12.27$; 177 men, 169 women, 2 unspecified).

Procedure

The study consisted of three parts. In the first part, each participant was randomly assigned to learn that either a man or a woman had communicated with air traffic control during a flight. Participants provided their priors, posteriors, and likelihoods for this scenario, just as they did for the doctor scenario in Study 4. As before,

a model posterior was computed for each participant and compared with his or her reported posterior. In the second part, participants completed filler tasks consisting of unrelated statistical judgments (e.g., "What percentage of the earth's surface is covered by land?") and trivia (e.g., "The German word *kummerspeck* means excess weight gained from emotional overeating"). In the third part, participants completed the same procedure as in Study 1, in which they indicated which of three statements they agreed with and evaluated person *X*, who made the Bayesian judgment that a man who performed surgery is more likely to be a doctor than a woman who performed surgery.

Results

Bayesian judgments were again observed, which replicates previous results (see Fig. S14 in the Supplemental Material). Model posteriors favored the man over the woman to be the pilot, even though both targets had communicated with air traffic control during a flight, man: $M = 94.0\%$, woman: $M = 63.1\%$, $b = 0.31$, $t(346) = 13.53$, $p < .0001$, $r = .59$, 95% CI = [.52, .62]. As before, this disparity was also observed among participants' reported posteriors, man: $M = 89.7\%$, woman: $M = 67.8\%$, $b = 0.22$, $t(346) = 9.56$, $p < .0001$, $r = .46$, 95% CI = [.37, .50].

Relatively small differences were observed between model posteriors and reported posteriors among participants who learned that the man had communicated with air traffic control, model posterior: $M = 94.0\%$, reported posterior: $M = 89.7\%$, $b = 0.04$, $t(692) = 2.43$, $p = .02$, $r = .09$, 95% CI = [.05, .15], and among participants who had learned that the woman had communicated with air traffic control, model posterior: $M = 63.1\%$, reported posterior: $M = 67.8\%$, $b = -0.05$, $t(692) = -2.64$, $p = .008$, $r = .10$, 95% CI = [.01, .21], further replicating previous results. So once again, posteriors reported by participants were close to the posteriors they should have reported, according to Bayesian rationality.

These very same participants who made Bayesian judgments agreed with the egalitarian judgment in a conceptually identical problem, albeit the effect was weakened because base rates were made salient by the first parts of the procedure. Seventy-nine percent of participants agreed that a man and a woman are equally likely to be a doctor given that they both performed surgery, 20% agreed that a man is more likely to be a doctor, and 1% agreed that a woman is more likely to be a doctor. Additional analyses (reported in the Supplemental Material) reveal the percentage of participants (71%) who used gendered base rates when making their probability judgments but not when indicating which judgment they agreed with (see Table S5 in the Supplemental Material).

Person *X*, who made a Bayesian judgment like participants did, was seen as unfair, $M = 3.12$, $SE = 0.09$; unjust, $M = 3.24$, $SE = 0.09$; inaccurate, $M = 3.42$, $SE = 0.10$; and unintelligent, $M = 3.38$, $SE = 0.08$, as indicated by means below the midpoint of 4 on the scales, Cronbach's $\alpha = .91$, composite $M = 3.29$, $SE = 0.08$, one-sample $t(347) = -8.90$, $p < .0001$, Cohen's $d = 0.48$, 95% CI = [0.37, 0.60]. Thus, the very same participants who criticized person *X*'s morality and intellect had just made judgments that were conceptually identical to person *X*'s Bayesian judgment.

However, the critical test of this inconsistency concerns the relationship between participants' evaluations of person *X* and their reported probability that the man versus the woman is the pilot conditional on having communicated with air traffic control. Perhaps participants who criticized person *X* also made egalitarian judgments that give the man and the woman equal probabilities of being the pilot.

But as shown in Figure 4, participants judged that the man was more likely than the woman to be the pilot, regardless of their evaluation of person *X*, $F(1, 344) = 84.58$, $p < .0001$, $\eta^2 = .19$, 95% CI = [.13, .27]. Even participants who were the most critical of person *X*—those who gave ratings of 1 on all four Likert-type items—judged that the man was more likely to be the pilot than the woman, fitted reported posteriors: man = 91.6%, woman = 81.4%; $b = -0.10$, $SE = 0.04$, $t(344) = -2.27$, $p = .02$, $r = .12$, 95% CI = [.03, .21].

The difference in probability judgments of the male and female targets increased as evaluations of person *X* became more positive, $F(1, 344) = 10.71$, $p = .001$, $\eta^2 = .02$, 95% CI = [.005, .07]. However, participants were equally and highly accurate irrespective of how they felt toward person *X*, as evidenced by the minimal difference between their model and reported posteriors across the entire range of evaluations (see Fig. S15 in the Supplemental Material). Thus, participants accurately judged that the man was more likely than the woman to be the pilot; these participants then criticized person *X* for making a conceptually similar Bayesian judgment. A study in the Supplemental Material replicated this key analysis when Bayesian judgments were elicited through the doctor scenario and person *X*'s statement concerned who is more likely to be the pilot (see Figs. S16–S18 in the Supplemental Material).

General Discussion

When presented with a third party who made a Bayesian judgment, participants criticized the morality and intellect of this person, shared less money with this person, and incurred financial costs on themselves to punish this person. However, participants made the

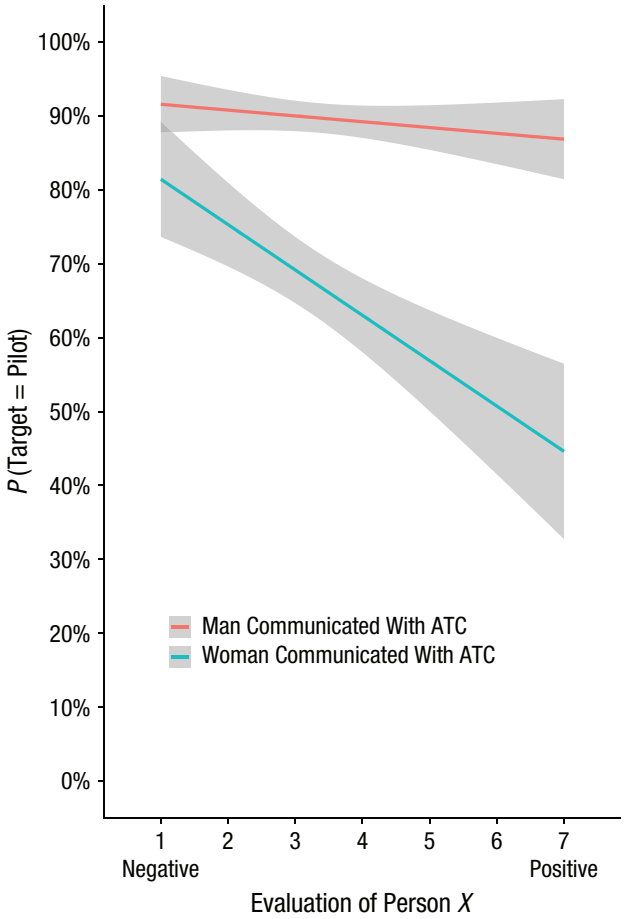

Fig. 4. Results from Study 5: reported posterior probabilities as a function of evaluations of person *X* (average of four Likert-type items). Gray bands indicate standard errors. ATC = air traffic control.

same judgment they criticized someone else for making, and they did so as a Bayesian statistician would.

Although statistical judgments typically lack moral flavor, it appears that under some circumstances—such as when the profession is doctor but not butcher, firefighter, or construction worker—these judgments are perceived as immoral, despite their accuracy. This finding dovetails with the work by Tetlock, Kristel, Elson, Green, and Lerner (2000), who coined the term *forbidden base rates* to refer to statistics that some may find offensive but nevertheless maximize accuracy. Financially incentivizing accuracy may increase the rate at which participants accept forbidden base rates. Future research may establish the demand curve for expressing accurate positions that are deemed unfair.

Previous work pitting statistics and morality against each other has relied on juxtaposing explicit and implicit measures (e.g., Cao & Banaji, 2016). But here, only explicit measures were used; participants faced no time pressure and were free to exercise full control over their responses. So, although participants' evaluations of person *X* and their own statistical judgments could have aligned, there was an inconsistency between these two sets of findings.

This inconsistency can be resolved if negative evaluations of person *X* were also the result of Bayesian inference. Both an unabashed sexist and a feminist statistician can state that a man who performed surgery is more likely to be a doctor than is a woman who performed surgery, albeit for different reasons. Given this uncertainty, participants may have attended to the base rate that person *X* is, a priori, more likely to be a sexist than a statistician. Insofar as participants correctly integrated this base rate with likelihood estimates of the probability that a sexist versus a statistician would say what person *X* said, criticisms of person *X* would also be Bayesian. Given that participants' probability judgments were Bayesian by taking into account the base rate that a doctor is, a priori, more likely to be a man than a woman, it is possible that their evaluations of person *X* were as well. In this case, participants did not exhibit hypocrisy by making the same judgment that they found repugnant when made by someone else. Rather, differences in which base rate is attended to would account for the observed inconsistency. Testing this possibility would build on efforts to formalize the process by which the motivation to assess character relates to probability judgments (Kleiman-Weiner, Shaw, & Tenenbaum, 2017; Pizarro & Tannenbaum, 2012).

But even if negative evaluations of person *X* were the result of Bayesian inference, the output of this inferential process was still at odds with participants' own statistical judgments. By condemning person *X* for favoring a man over a woman to be a doctor, both via negative ratings of morality and intellect and via financial decisions, participants exerted their desire for equal judgments of the man and woman, a position they undercut by judging that the man is more likely than the woman to be the doctor. It may be the case, then, that the same cognitive process of Bayesian inference underlies statistical judgments and negative evaluations of other people who make certain statistical judgments.

One limitation here is that data were collected from MTurk, so participants may have been inattentive. However, this venue is an appropriate place to demonstrate these effects because it is on the Internet where people commonly express outrage at violations of egalitarian norms (Crockett, 2017). Furthermore, results from MTurk are comparable with results obtained from laboratory settings (Amir, Rand, & Gal, 2012).

Finally, these findings have implications for criminal trials in which it is illegal to use group membership to assess guilt (Kohler, 1992). Even if this law is endorsed, priors based on group membership may influence the mental computations of judges and jurors. Although

this may be Bayesian, it may also result in unequal judgments that the law is designed to prevent, thereby further compounding inequalities (Loury, 2002). Thus, people's own statistical savvy may be a barrier to the equal treatment they desire.

Action Editor

Ayelet Fishbach served as action editor for this article.

Author Contributions

All the authors designed the research. J. Cao performed the research and analyzed the data under the supervision of M. R. Banaji. J. Cao drafted the manuscript, and M. Kleiman-Weiner and M. R. Banaji provided critical revisions. All the authors approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797618805750

Open Practices

All data and materials have been made publicly available via the Open Science Framework and can be accessed at osf.io/9fdhb. The design and analysis plans for these studies were not preregistered. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797618805750. This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

Note

1. The term *judgment* is used to refer to how likely a man, as opposed to a woman, is to be a doctor, conditional on both people having performed surgery. Participants also made judgments of someone else who offered the Bayesian judgment; these data are discussed using the term *evaluation*.

References

Abelson, R. P. (1968). Psychological implication. In R. P. Abelson, E. Aronson, W. J. McGuire, T. M. Newcomb, M. J. Rosenberg, & P. H. Tannenbaum (Eds.), *Theories of cognitive consistency: A sourcebook* (pp. 112–139). Chicago, IL: Rand McNally.

Amir, O., Rand, D. G., & Gal, Y. K. (2012). Economic games on the Internet: The effect of $1 stakes. *PLOS ONE*, *7*(12), Article e31461. doi:10.1371/journal.pone.0031461

Auspurg, K., Hinz, T., & Sauer, C. (2017). Why should women get less? Evidence on the gender pay gap from multifactorial survey experiments. *American Sociological Review*, *82*, 179–210.

Biernat, M., & Kobrynowicz, D. (1997). Gender- and race-based standards of competence: Lower minimum standards but higher ability standards for devalued groups. *Journal of Personality and Social Psychology*, *72*, 544–557.

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

Cao, J., & Banaji, M. R. (2016). The base rate principle and the fairness principle in social judgment. *Proceedings of the National Academy of Sciences, USA*, *113*, 7475–7480.

Cao, J., Kleiman-Weiner, M., & Banaji, M. R. (2017). Statistically inaccurate and morally unfair judgments via base rate intrusion. *Nature Human Behaviour*, *110*, 738–742.

Constitute. (2016). *Equality regardless of gender*. Retrieved from https://constituteproject.org/ontology/equalgr1/Minority_Rights/rights_and_duties?lang=en

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*, 769–771.

Dworkin, R. (2002). *Sovereign virtue: The theory and practice of equality*. Cambridge, MA: Harvard University Press.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*, 573–598.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge, England: Cambridge University Press.

Foschi, M. (1996). Double standards in the evaluation of men and women. *Social Psychology Quarterly*, *59*, 237–254.

Garnham, A., Doehren, S., & Gygax, P. (2015). True gender ratios and stereotype rating norms. *Frontiers in Psychology*, *6*, Article 1023. doi:10.3389/fpsyg.2015.01023

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*, 767–773.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*, 430–454.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–251.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*, 271–304.

Kleiman-Weiner, M., Shaw, A., & Tenenbaum, J. B. (2017). Constructing social preferences from anticipated judgments: When impartial inequity is fair and why? In *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 676–681). Oakbrook Terrace, IL: Cognitive Science Society.

Kohler, J. J. (1992). Probabilities in the courtroom: An evaluation of the objections and policies. In D. K. Kagehiro & W. S. Laufer (Eds.), *Handbook of psychology and law* (pp. 167–184). New York, NY: Springer.

Loury, G. C. (2002). *The anatomy of racial inequality*. Cambridge, MA: Harvard University Press.

McAuliffe, K., Blake, P. R., Steinbeis, N., & Warneken, F. (2017). The developmental foundations of human fairness. *Nature Human Behaviour*, *1*, Article 0042. doi:10.1038/s41562-016-0042

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences, USA*, *109*, 16474–16479.

Oliver, L. (2017, August 30). Meet the nurse who will soon perform surgery on patients alone. *The Guardian*. Retrieved from https://www.theguardian.com/healthcare-network/2017/aug/30/nurse-perform-surgery-patients-alone-surgical-care-practitioner

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.

Pizarro, D. A., & Tannenbaum, D. (2012). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 91–108). Washington, DC: American Psychological Association.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*, 811–832.

Price Waterhouse v. Hopkins. 490 U.S. 228 (1989).

Rawls, J. (2001). *Justice as fairness: A restatement*. Cambridge, MA: Harvard University Press.

Rocheleau, M. (2017, March 7). Chart: The percentage of women and men in each profession. *Boston Globe*. Retrieved from https://www.bostonglobe.com/metro/2017/03/06/chart-the-percentage-women-and-men-each-profession/GBX22YsWl0XaeHghwXfE4H/story.html

Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues*, *57*, 743–762.

Shaw, A. (2016). Fairness: What it isn't, what it is, and what it might be for. In D. C. Geary & D. B. Berch (Eds.), *Evolutionary perspectives on child development and education* (pp. 193–214). Basel, Switzerland: Springer International.

Shaw, A., & Olson, K. (2014). Fairness as partiality aversion: The development of procedural justice. *Journal of Experimental Child Psychology*, *119*, 40–53.

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, *78*, 853–870.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245–272.