Bayesian Reasoning the Social Domain

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th>Citation</th>
<th>Cao, Jack. 2019. Bayesian Reasoning the Social Domain. Doctoral dissertation, Harvard University, Graduate School of Arts &amp; Sciences.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:42029778">http://nrs.harvard.edu/urn-3:HUL.InstRepos:42029778</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Bayesian Reasoning in the Social Domain

A dissertation presented

by

Jack Cao

to

The Department of Psychology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Psychology

Harvard University

Cambridge, Massachusetts

March 4, 2019
© 2019 Jack Cao

All rights reserved.
Bayesian Reasoning in the Social Domain

ABSTRACT

People are simultaneously attuned to group-based statistics and to norms concerning fairness. While group-based statistics support Bayesian principles, norms concerning fairness underlie egalitarian values. This dissertation examines the judgments people make when Bayesian principles and egalitarian values are both at stake. The findings of three papers reveal that people employ group-based statistics in their social judgments even when they believe it is inappropriate to do so. In Paper 1, people eschewed obvious and relevant statistics in their explicit judgments but hewed with these statistics in their implicit judgments – both before and after individuating information was learned. In Paper 2, people shared less money with a third party who made a judgment that was consistent with established statistics. These people even incurred financial costs on themselves to punish this third party. However, these very same people used statistics as a Bayesian statistician would to render the same judgment that they found morally and intellectually repugnant when offered by someone else. In Paper 3, people continued to rely on statistics even when Bayesian reasoning dictated that these statistics should be ignored. This resulted in not only a statistical error, but also a moral error according to people’s own standards of conduct. These findings suggest that people’s unwitting reliance on statistics may be a barrier to the fairness they desire. This dissertation concludes with a discussion of how psychological science can contribute to newly emerging but pressing issues about the rise of artificial intelligence in what were once exclusively human domains.
TABLE OF CONTENTS

ACKNOWLEDGEMENTS .................................................................................................................. v
INTRODUCTION ................................................................................................................................. 1

PAPER 1: Cao & Banaji (2016), *Proc Natl Acad Sci* ................................................................. 6
  Abstract ........................................................................................................................................ 6
  Significance Statement and Introduction .................................................................................. 7
  Experiment 1 ................................................................................................................................. 13
  Experiment 2 ................................................................................................................................. 17
  Experiment 3 ................................................................................................................................. 19
  Experiment 4 ................................................................................................................................. 21
  Discussion .................................................................................................................................. 24
  Methods ........................................................................................................................................ 29

  Abstract .......................................................................................................................................... 31
  Introduction .................................................................................................................................... 32
  Experiment 1 .................................................................................................................................. 36
  Experiment 2 .................................................................................................................................. 38
  Experiment 3 .................................................................................................................................. 41
  Experiment 4 .................................................................................................................................. 45
  Experiment 5 .................................................................................................................................... 52
  Discussion .................................................................................................................................... 57

PAPER 3: Cao, Kleiman-Weiner, & Banaji (2017), *Nat Hum Behav* ...................................... 60
  Abstract .......................................................................................................................................... 60
  Introduction .................................................................................................................................... 61
  Experiment 1 .................................................................................................................................. 66
  Experiment 2 .................................................................................................................................. 68
  Experiment 3 .................................................................................................................................. 69
  Experiment 4 .................................................................................................................................. 70
  Discussion .................................................................................................................................... 73
  Methods ........................................................................................................................................ 77

GENERAL DISCUSSION .................................................................................................................... 79
REFERENCES ..................................................................................................................................... 94
PAPER 1 SUPPLEMENTAL INFORMATION .................................................................................. 104
PAPER 2 SUPPLEMENTAL INFORMATION .................................................................................. 118
PAPER 3 SUPPLEMENTAL INFORMATION .................................................................................. 144

---

1 This journal does not include section breaks in its articles. Section breaks have been added here for clarity.
ACKNOWLEDGEMENTS

For being on my committee, I thank Sam Gershman, Fiery Cushman, Francesa Gino, and Mike Norton. Your work has inspired new ways of thinking that keep young scholars like myself busy and hungry. Thank you for shaping the ideas in this dissertation.

For building and protecting an environment where ideas are treasured, I thank Harvard University.

For challenging me to make my ideas worthy of financial support, I thank the National Science Foundation Graduate Research Fellowship Program; Inequality and Social Policy Program at Harvard Kennedy School of Government; and Harvard’s Mind, Brain, and Behavior Initiative.

For their research assistance that evolved from bringing papers to publication to seeding their own ideas that have begun to take on lives of their own, I thank Kirsten Morehouse, Juan Lopez Martin, Natalia Dashan, and Christopher Willis.

For their tireless skepticism and tough questioning that have improved in countless ways how I present data and frame arguments, I thank the Social Cognition Laboratory.

For guiding me through an intellectual world different from my own, giving me the joy of sharing social psychology, and working together to make a contribution, I thank Max Kleiman-Weiner.

For our many impromptu conversations about math, statistics, and human intuition, I thank my brother, Alex.

For nurturing in me and my brother the curiosity that made all those impromptu conversations possible and enjoyable, I thank my parents, June and Jim.
For giving me everyday at least one reason to laugh and for volunteering as perennial first pilot subject, I thank my partner, Kate.

Finally, I thank my advisor, Mahzarin R. Banaji, who showed me the wonder in changing one’s own mind in the face of evidence. Much of my own mind has changed for the better because of her.
INTRODUCTION

To make the best judgments, we need to consider new information. Upon learning that a patient has begun to exhibit a different set of symptoms, a doctor might reformulate her diagnosis. After reading students’ course evaluations, a professor may change her perspective on how the semester progressed. And as readers scrutinize this dissertation, they might update their views on whether its writer is worthy of a Harvard PhD.

But exactly how should new information be incorporated into the updating process? An 18th century English Presbyterian minister named Thomas Bayes answered this question with a theorem that now bears his name. Derived from the axioms of probability theory, Bayes’ theorem captures the intuition that the signal strength of new information matters a great deal for changing one’s judgments. To illustrate, a thorough, firsthand examination of a patient provides a stronger basis for a different diagnosis than a brief telephone consultation. Whereas the former provides a doctor with a strong signal of what the underlying disease might be, the latter is noisy and unreliable.

Critically, Bayes’ theorem states that the signal strength of new information cannot be considered in isolation; it must be integrated with prior beliefs. Imagine a doctor who learns that her patient suffers from coughing, chest pains, and wheezing. Although these symptoms could indicate that the patient has lung cancer, it is more likely that the patient has the flu. Bayes’ theorem formalizes why the flu is indeed more likely than lung cancer when the symptoms are consistent with both illnesses. People are more likely to contract the flu than develop lung cancer. This base rate – or statistical regularity – forms the basis of prior beliefs
that guide judgments to what is most probable. Insofar as these priors are set aside, Bayesian reasoning suffers.

In the field of judgment and decision making, a rich tradition of comparing human judgments with Bayesian prescriptions has revealed that while people attend to the signal strength of new information, they tend to ignore base rates (Kahneman & Tversky, 1973; Bar-Hillel 1980). Base rate neglect, as this error is known, not only undermines the rationality of human judgments in laboratory studies, but can also lead to poor decisions with serious consequences.

Medical professionals became aware of this through the mammogram problem: a woman gets positive mammogram during a routine screening, and 2% of women have breast cancer. How likely is this particular woman to have breast cancer? Because doctors did not fully consider the base rate of 2%, they overestimated how likely this particular woman was to have breast cancer (Eddy, 1982). At scale, these overestimations have been linked to testing that is unnecessary, expensive, and apprehension inducing for patients and care providers (Bleyer & Welch, 2012).

Medical professionals have since heeded to Reverend Bayes, and they are not alone. Authors of popular press books aimed at improving how people live are sure to include a chapter on Bayesian reasoning (Silver, 2012; Duke, 2018). Programmers who create spam filters for e-mail accounts rely on Bayes’ theorem to calculate how likely an incoming message is to be spam (Sahami, Dumais, Heckerman, & Horvitz, 1998). And in response to psychology’s replication crisis, a growing chorus has advocated for Bayesian approaches to data analysis (Lee & Wagenmakers, 2013; Trafimow & Marks 2015; Colquhuon, 2017). Small wonder, then, that
Reverend Bayes’ work has been praised by the statistician Jim Berger (2000) as “arguably the most powerful mechanism created for processing data and knowledge” and described by the science writer Sharon Beth McGrayne (2012) as “the theory that would not die.” More than not dying, it appears that Bayesian theory is flourishing.

Despite the reach and impact of Bayes’ theorem, its influence on social judgments has elicited skepticism and controversy. The problem arises from the prescription that judgments be rooted not only in new information about an individual, but also base rates about the social group to which the individual belongs. In social psychology, decades of research have characterized base rates about social groups as stereotypes (Locksley, Hepburn, & Ortiz, 1982; Krosnick, Li, & Lehman, 1990; Jussim, 2012) and documented the harms that result from stereotyping – harms such as intergroup prejudice and discrimination (Allport, 1954) and insidious double standards (Biernat, 2002). Given that people have systematically changed over time in endorsement of greater beliefs in egalitarianism – as evidenced by laboratory studies (Gaertner & Dovidio, 1986) and analyses of large datasets (Charlesworth & Banaji, 2019) – eschewing stereotypes, even if their use is statistically justified by Bayes’ theorem, would seem prudent.

In lieu of base rates as stereotypes, many theories of morality privilege only information that is specific to the individual (Dworkin, 2000; Rawls, 2001). This position, in fact, undergirds legal requirements that base rates be set aside when determining guilt in American courtrooms (Kohler, 1992) and insurance premiums in the European Union (Test Achats v. Council of Ministers, 2011). While these legal requirements technically violate Bayesian principles, they uphold moral principles that promote the fair treatment of individuals.
Unlike the law where the question of statistics versus morality has been clearly settled in favor of morality, the new field of artificial intelligence is only beginning to grapple with this dilemma. Take Google Translate, for instance. The Turkish sentences “O bir doktor” and “O bir hemsire” mean, respectively, “One is a doctor” and “One is a nurse.” The pronoun O is gender neutral and both sentences clearly begin with this pronoun. But before late 2018, these sentences were translated to, respectively, “He is a doctor” and “She is a nurse.” Google Translate, in some way, took into account the base rate that doctors tend to be men and nurses women. But in an effort to promote fairness, the translation system has since been revised to output both masculine and feminine versions of gender neutral words (Google, 2018).

Curiously, this revision only appears to have effect on single sentence inputs. For example, O bir doktor is translated to both He is a doctor and She is a doctor, as intended. But when multiple sentences are entered as input, the output is still gender biased: “O bir doktor. O bir hemsire.” is translated to “He is a doctor. She is a nurse.” – just as it was before. This behavior by the translation system appears inconsistent with the goal of promoting fairness. Google has been made aware of this issue, and it remains to be seen what additional changes, if any, will be made.

So clearly, the question of how to handle base rates about social groups is relevant to developers of artificial intelligence (Caliskan, Bryson, & Narayanan, 2017). However, this question is one that that concerns everyday people who are attuned to these statistics (Garnham, Doehren, & Gygax, 2015) and to values promoting fairness (Shaw, 2016). This dissertation examines the judgments people make when Bayesian principles and egalitarian values are both at stake. Do people prefer to make judgments that uphold either the statistical
imperative to consider base rates or the moral imperative to set those base rates aside? What are the behavioral implications of this preference? And to what extent are their actual judgments consistent with these preferences?

This work is situated at the intersection of judgment and decision making and moral psychology. Bayesian models are common in judgment and decision making research, but rarely does this work incorporate the moral concerns that pervade when considering the individual and the individual’s social group. And while moral psychology examines how people think about issues of right and wrong, prescriptions of how base rates ought to influence judgments typically are not part of the fray. It is at the intersection of these two fields where one can learn about how people navigate the tension between Bayesian principles and egalitarian values.

To investigate this tension within the human mind, this dissertation employs a wide range of methods. Paper 1, published in *Proceedings of the National Academy of Sciences*, juxtaposes explicit and implicit measures to test the consistency between judgments that are made under full conscious control and those that are not. Paper 2, published in *Psychological Science*, employs economic games to assess the strength of people’s conviction regarding the use of base rates in social judgments. Paper 3, published in *Nature Human Behaviour*, capitalizes on Bayesian networks to provide people with the opportunity to make judgments that are consistent with both the statistical and moral imperatives – instead of just one at the expense of the other. Below are verbatim reproductions of these three papers, followed by an as of yet unpublished discussion that focuses on what psychological science might contribute to newly emerging but pressing issues about the rise of algorithms and artificial intelligence in what were once exclusively human domains.
Abstract

Meet Jonathan and Elizabeth. One person is a doctor and the other is a nurse. Who is the doctor? When nothing else is known, the base rate principle favors Jonathan to be the doctor and the fairness principle favors both individuals equally. But when individuating facts reveal who is actually the doctor, base rates and fairness become irrelevant, as the facts make the correct answer clear. In three experiments, explicit and implicit beliefs were measured before and after individuating facts were learned. These facts were either stereotypic (e.g., Jonathan is the doctor, Elizabeth is the nurse) or counterstereotypic (e.g., Elizabeth is the doctor, Jonathan is the nurse). Results showed that before individuating facts were learned, explicit beliefs followed the fairness principle whereas implicit beliefs followed the base rate principle. After individuating facts were learned, explicit beliefs correctly aligned with stereotypic and counterstereotypic facts. Implicit beliefs, however, were immune to counterstereotypic facts and continued to follow the base rate principle. Having established the robustness and generality of these results, a fourth experiment verified that gender stereotypes played a causal role: when both individuals were male, explicit and implicit beliefs alike correctly converged with individuating facts. Together, these experiments demonstrate that explicit beliefs uphold fairness and incorporate obvious and relevant facts, but implicit beliefs uphold base rates and appear impervious to counterstereotypic facts.
Significance Statement

In the absence of individuating facts, beliefs about individuals can either employ base rates to maximize statistical likelihood or uphold fairness to maximize equal opportunity. But once individuating facts become known, neither base rates nor fairness should drive beliefs. Only the facts should matter. While explicit beliefs rationally follow this prescription, implicit beliefs do not. Despite learning individuating facts about a particular male and female that rendered base rates inapplicable, implicit beliefs still relied on base rates. These findings are important not just for theories of social cognition and Bayesian updating, but also for crafting policies that will minimize the undesired impact of stereotypes on decisions about the worth and capabilities of specific individuals.

Introduction

Imagine meeting Jonathan and Elizabeth. One person is a doctor. The other is a nurse. Who is the doctor? Or imagine that an employer is deciding to hire either Colin or Jamaal. A background check will reveal that one person has a violent felony on his record and therefore will not be hired. Who is the violent felon? Before individuating facts are learned, when only gender or race is known, one of two principles can guide beliefs.

The first, which we call the *base rate principle*, supports the belief that Jonathan is the doctor and Jamaal is the violent felon. If ignoring base rates is considered an error, then one must realize that doctors are more likely to be men than women (Kaiser Family Foundation, 2015) and people with violent felonies on their record are more likely to be Black than White (Bureau of Justice Statistics, 2015). In fact, since group membership contains useful information
for deciding whether an individual has a certain attribute, stereotypes have been conceptualized as base rates (Locksley et al., 1980; Locksley et al., 1982; Krosnick et al., 1990; Jussim, 2012). Moreover, decision theorists have shown that base rates are critical ingredients for making predictions (Beyth-Marom & Fischhoff, 1983; Bar-Hillel, 1980), as neglecting base rates will cause predictions to deviate from what is statistically likely (Tversky & Kahneman, 1974).

Using these base rates, however, is inconsistent with a second principle that we call the fairness principle. By this account, it is morally proper to assume a fair coin, so to speak. Jonathan and Elizabeth are equally likely to be the doctor and Colin and Jamaal are equally likely to have a violent felony on their record. Motivated by egalitarian values, many people believe that base rates cannot and should not be used to make such predictions. In fact, the value of fairness is deeply woven into many legal systems. American courts have rejected the use of base rates to determine guilt (Kohler, 1982; Schauer, 2003), and the European Union has banned gender-based insurance premiums (Test-Achats v. Council of Ministers, 2011).

In the present work, we assess which principle guides beliefs before individuating facts are learned. Given only information about gender, do beliefs favor Jonathan to be the doctor or both Jonathan and Elizabeth equally to be the doctor? We then assess if the base rate and fairness principles are set aside after individuating facts are learned. Given facts that make abundantly clear who is – and who is not – the doctor, do beliefs align with the facts?

In Experiment 1, participants meet Jonathan and Elizabeth and must predict who is the doctor and who is the nurse. If participants employ base rates, then Jonathan will be more likely than Elizabeth to be the doctor. But if participants privilege fairness, then both Jonathan
and Elizabeth will be equally likely to be the doctor.\footnote{We do not imply that the base rate and fairness principles are mutually exclusive. Our measures allow participants to employ both principles in their initial beliefs.} Next, participants were taught one of three types of individuating facts: 1) stereotypic facts: Jonathan is the doctor; Elizabeth is the nurse, 2) counterstereotypic facts: Elizabeth is the doctor; Jonathan is the nurse, or 3) irrelevant facts that served as a control: Jonathan vacationed in Colorado; Elizabeth vacationed in California. Finally, participants indicated their beliefs about each individual’s profession once again.

Before participants learn individuating facts, beliefs about Jonathan and Elizabeth’s professions can follow either the base rate principle or the fairness principle. But after participants learn individuating facts, beliefs should follow neither principle. At this point, when the task is simply to restate what the facts have made plainly obvious, Jonathan and Elizabeth are no longer stand-ins for male and female, respectively; they have become individuated (Fiske & Neuberg, 2013). As such, beliefs about them should align with clear-cut facts instead of with broad principles that no longer apply to these individuals now that their actual professions are known.

Before and after learning individuating facts, participants indicated their explicit beliefs on a Likert-type scale and their implicit beliefs on an Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), which measured the strength of association between each individual – Jonathan vs. Elizabeth – and the central attribute of doctor vs. nurse. There is a wealth of evidence showing that explicit and implicit responses jointly and uniquely predict behavior (Cameron et al., 2012; Greenwald et al., 2009) even though they can be dissociated (Nosek et al., 2007).
Administering a Likert-type scale after teaching participants Jonathan and Elizabeth’s professions amounts to little more than a manipulation check. So regardless of whether the individuating facts are stereotypic or counterstereotypic, explicit beliefs reported afterwards should align with the facts if these beliefs are to be considered appropriate. To assess the appropriateness of implicit beliefs about Jonathan and Elizabeth’s professions once individuating facts are learned, we rely on participants’ own explicit beliefs as the normative standard. If participants’ implicit beliefs are inconsistent with their explicit beliefs – which will likely reflect each individual’s actual profession given the clarity of the facts and the simplicity of the task – then such implicit beliefs would be inappropriate, for these beliefs would contradict both the facts and what participants themselves identify as correct. Therefore, if Jonathan is the doctor and Elizabeth is the nurse and if participants explicitly agree, then Jonathan should be associated with *doctor* on the IAT. But if Elizabeth is the doctor and Jonathan is the nurse and if participants explicitly agree, then Elizabeth should be associated with *doctor* on the IAT.

Some research suggests that implicit associations are less malleable than their explicit counterparts (Gregg, Seibt, & Banaji, 2006; Rydell et al., 2007), making it seem unlikely that implicit beliefs will incorporate individuating facts like explicit beliefs would. 2 However, other research has identified conditions under which implicit associations appear highly amenable to new information (Blair, 2002; Blair, Ma, & Lenton, 2001; Wyer, 2016; Wyer, 2010; Cone &

---

2 We use the term *implicit belief* since implicit associations between individual and profession are truth evaluable when compared to the individuating facts and to participants’ own explicit beliefs. If the facts are stereotypic, then an association between Jonathan and doctor would correctly reflect the facts. If the facts are counterstereotypic, then an association between Elizabeth and doctor would correctly reflect the facts. There is a lively debate on whether implicit associations are propositional vs. associative in nature (Gawronksi & Bodenhausen, 2006; Mitchell, De Houwer, Lovibond, 2009). The use of the term *implicit belief* should not be interpreted as a position on this debate, which this paper does not address.
Ferguson, 2015). Drawing from this latter research, we have incorporated three aspects into the experimental paradigm that together create favorable conditions for implicit beliefs to align with the facts. Given these favorable conditions, it would be surprising if implicit beliefs still did not reflect individuating facts.

First, consistent with work demonstrating that highly diagnostic information can shift implicit evaluations (Cone & Ferguson, 2015), we provide information that is the most diagnostic of Jonathan and Elizabeth’s professions: these individuals’ actual professions. In addition to maximizing the diagnosticity of the facts, we also minimize the standard for what is considered a correct implicit belief given the individuating facts. IAT $D$ scores, which are taken to reflect implicit beliefs, should be on the correct side of a neutral score of zero. If Elizabeth is the doctor, she need not to be as strongly associated with doctor compared to when Jonathan is the doctor. Nonetheless, she should still be associated with doctor, which is her actual profession when the facts are counterstereotypic.

Second, we test the updating of mental representations of specific individuals. Through the unambiguous facts we teach, we construct representations of these individuals where any and all variability is removed: the doctor is Jonathan and the nurse is Elizabeth, or vice versa. This focus on the individual differs from most work in social cognition that has sought to update mental representations of entire social groups (Gregg et al., 2006; Wyer, 2016; Lai et al., 2014). Groups contain variability because the distribution of an attribute (e.g., doctor vs. nurse) across a group (e.g., male vs. female) is broad: there are doctors and nurses of both genders. It is for this reason that on the IAT, Jonathan and Elizabeth are used instead of Male and Female. The association between gender and profession may not change drastically in response to
individuating facts. But given the watertight certainty that is possible to obtain when considering individuals instead of groups, the association between specific individuals and profession should correctly give way to individuating facts.

Third, instead of examining preferences, evaluations, or attitudes – which are all inherently subjective – we examine fact-based beliefs. Even young children know that if person A likes red and person B likes green, both A and B can be right. But these children also know that facts hold a different status: if A thinks germs are big and B thinks germs are small, only one person can be right (Heiphetz et al., 2013). After learning individuating facts, there can only be one correct belief for a logical learner. If the facts are stereotypic, then we must believe, explicitly and implicitly, that Jonathan is the doctor. If the facts are counterstereotypic, then we must believe, explicitly and implicitly, that Elizabeth is the doctor. So will updated beliefs about Jonathan and Elizabeth’s professions reflect clear-cut individuating facts? Or will these beliefs still contain traces of the base rate principle or fairness principle?

---

3 Some readers may note the relevance of Bayesian models. While beliefs are measured before and after new facts are presented, these beliefs are not priors and posteriors per se because they are not probability estimates. Moreover, the lack of any uncertainty in the facts we present undercuts the need to measure likelihoods, which are necessary for the Bayesian analyses that related research has undertaken (Locksley et al., 1980). Therefore, the data presented are not suited for a formal Bayesian analysis and we do not make any strong claims about whether belief updating follows the prescriptions of Bayes’ theorem.
Experiment 1

Before the facts. Before learning Jonathan and Elizabeth’s actual professions, participants (N = 574) reported explicit beliefs on a Likert-type scale that favored Jonathan to be the doctor and Elizabeth to be the nurse [M = -0.30, one-sample t(573) = -11.74, p < .0001, Cohen’s d = 0.49]. Implicit beliefs, measured using IAT D scores, also favored Jonathan to be the doctor and Elizabeth to be the nurse, although to a much greater extent [M = -0.43, one-sample t(573) = -27.23, p < .0001, Cohen’s d = 1.14].

Both explicit and implicit beliefs were, on average, consistent with base rate usage. However, visual inspection of each distribution reveals a stark difference (Fig. S1). Whereas the overwhelming majority of participants explicitly agreed with a statement consistent with the fairness principle, an overwhelming majority of the same participants displayed implicit beliefs consistent with the base rate principle.

After the facts. After Jonathan and Elizabeth’s professions were presented, we administered the same Likert-type scale and IAT again. Explicit beliefs in the control condition continued to adhere to the fairness principle [M_{after} = -0.14 vs. M_{before} = -0.24; p = .19]. In the experimental conditions, explicit beliefs were correctly updated to align with both stereotypic facts [M_{after} = -2.42 vs. M_{before} = -0.34; p < .0001] and counterstereotypic facts [M_{after} = 2.64 vs. M_{before} = -0.32; p < .0001]. Thus, when individuating facts made absolutely clear each person’s profession, explicit beliefs appropriately set aside the fairness principle (Fig. S2).
While explicit beliefs displayed a sensible pattern of updating, implicit beliefs aligned only with stereotypic facts. When individuating facts were counterstereotypic, implicit beliefs still utilized the base rate favoring Jonathan over Elizabeth to be the doctor (Fig. 1).

In the control condition, we expected and observed a standard test-retest effect\(^4\) such that implicit beliefs were closer to a neutral \(D\) score of zero \([M_{after} = -0.30 \text{ vs. } M_{before} = -0.43; p < 0.0001]\). Stereotypic facts had an effect above and beyond test-retest in the expected direction \([t(571) = -3.68, p = .0003]\), resulting in \(D\) scores consistent with each person’s actual profession \([M_{after} = -0.43, 95\% CI (-0.48, -0.38)]\).

Counterstereotypic facts also had an effect above and beyond test-retest in the expected direction \([t(571) = 2.57, p = .01]\). However, this effect was insufficient to bring implicit beliefs in line with the fact that Elizabeth, not Jonathan, is the doctor \([M_{after} = -0.20, 95\% CI (-0.25, -0.14)]\). Jonathan, who is actually the nurse, was more strongly associated with doctor while Elizabeth, who is actually the doctor, was more strongly associated with nurse.\(^5\) Despite the certainty provided by these individuating facts, implicit beliefs continued to rely on the base rate principle.

---

\(^4\) By taking the IAT twice, participants improve their ability sort stimuli quickly to the appropriate category, leading to reduced effects (Cunningham, Preacher, & Banaji, 2009).

\(^5\) Reactivity is possible in all pre-post designs, but the IAT does not appear to display the reactivity inherent in self-report measures. In fact, Lai and et al. (2014) used a Solomon four group design (i.e., random assignment to pre-test) to test the reactivity of the IAT and found little evidence for it. Thus, reactivity is not a concern here more so than in any other pre-post design.
In an effort to find evidence of correct updating, we examined only those participants \((n = 81)\) whose \(D\) scores on the first IAT were between -0.15 and 0.15, a range close to neutrality. We reasoned that without strong initial beliefs to temper the influence of contradicting facts, those participants who were randomly assigned to learn counterstereotypic facts might correctly associate *Elizabeth* with *doctor* and *Jonathan* with *nurse*.

But even in this subsample, implicit beliefs aligned only with stereotypic facts (Fig. 2). Participants who learned stereotypic facts produced \(D\) scores that reflected these facts \([M_{after} = -0.36, 95\% CI (-0.45, -0.28)]\), an effect that exceeded test-retest \(t(78) = -3.14, p = 0.002\).

However, participants who learned counterstereotypic facts failed to incorporate these facts \([M_{after} = -0.09, 95\% CI (-0.18, 0.004)]\). Remarkably, the effect of counterstereotypic facts did not
differ from test-retest \( t(78) = -0.13, p = .90 \). Even for participants who were on the cusp of aligning their implicit beliefs with counterstereotypic facts, implicit beliefs did not reflect these facts. The impact of these facts hardly differed from the impact of irrelevant control facts.

**Fig. 2** Experiment 1: Mean implicit beliefs about Jonathan and Elizabeth’s professions among only participants whose implicit beliefs before learning the facts were neutral \((n = 81)\). IAT \(D\) scores are on the \(y\)-axis. Error bars are 95% CIs.
Experiment 2

To ensure these findings were not specific to the idiosyncrasies of the names Jonathan and Elizabeth and the professions doctor and nurse, we replicated Experiment 1 with a new set of names and professions. In Experiment 2, we tested Richard and Jennifer and scientist and artist.

Before the facts. Before learning Richard and Jennifer’s actual professions, participants (N = 808) reported explicit beliefs that favored Richard to be the scientist and Jennifer to be the artist [M = -0.20, one-sample t(807) = -8.98, p < .0001, Cohen’s d = 0.32]. Implicit beliefs likewise favored Richard to be the scientist and Jennifer to be the artist, although to a much greater extent [M = -0.29, one-sample t(807) = -23.20, p < .0001, Cohen’s d = 0.82].

As before, visual inspection of each belief distribution reveals alignment with a different principle (Fig. S3). While most participants explicitly agreed with a statement consistent with the fairness principle, an overwhelming majority of the same participants displayed implicit beliefs consistent with the base rate principle.

After the facts. Once again, participants aligned their explicit beliefs with both stereotypic and counterstereotypic facts, thereby setting aside the fairness principle (Fig. S4). Upon learning stereotypic facts that Richard is the scientist and Jennifer is the artist, explicit beliefs reflected the facts [M after = -2.56 vs. M before = -0.13; p < .0001]. Upon learning counterstereotypic facts that Jennifer is the scientist and Richard is the artist, explicit beliefs reflected these facts [M after = 2.53 vs. M before = -0.21; p < .0001].
But as in Experiment 1, implicit beliefs aligned only with stereotypic facts. After learning counterstereotypic facts, implicit beliefs about Richard and Jennifer’s professions still utilized the base rate favoring Richard to be the scientist and Jennifer to be the artist (Fig. S5).

Stereotypic facts had an effect above and beyond test-retest in the expected direction \([t(805) = -2.98, p = .003]\), resulting in \(D\) scores consistent with Richard’s actual profession as the scientist and Jennifer’s actual profession as the artist \([M_{\text{after}} = -0.29, 95\% CI (-0.33, -0.25)]\).

Counterstereotypic facts also had an effect above and beyond test-retest in the expected direction \([t(805) = 4.03, p = .0001]\). However, this effect was insufficient to align implicit beliefs with the fact that Jennifer, not Richard, is the scientist \([M_{\text{after}} = -0.05, 95\% CI (-0.09, -0.007)]\). Implicit beliefs continued to rely on base rates that were no longer useful with respect to Richard and Jennifer.

Again, we analyzed the subsample of participants \((n = 172)\) whose \(D\) scores on the first IAT were in the neutral range between -0.15 and 0.15 (Fig. S6). Among these participants who learned stereotypic facts, implicit beliefs reflected Richard and Jennifer’s true professions \([M_{\text{after}} = -0.18, 95\% CI (-0.24, -0.12)]\). But among these participants who learned counterstereotypic facts, implicit beliefs once again failed to incorporate these facts \([M_{\text{after}} = -0.01, 95\% CI (-0.08, 0.05)]\). In this subsample, the effect of both stereotypic and counterstereotypic facts did not exceed test-retest \([t_{(169)} < |1.56|, ps > .12]\). Nonetheless, we again found that implicit beliefs align with stereotypic facts but not counterstereotypic facts, even among those participants whose initial implicit beliefs were neutral.
Experiment 3

Implicit beliefs might not have reflected counterstereotypic facts because participants may have regarded the targets not as specific individuals but rather as generic representatives of the many individuals who share these familiar names. Experiment 3 addresses this concern by using novel names – Lapper and Affina – which have not in prior experience been used to refer to any particular man or woman. If the data replicate those of Experiments 1 and 2, then we can conclude more confidently that the results reflect beliefs about specific individuals, as there are no others who share these novel names. Participants underwent the same procedure from Experiment 1 except they were initially told that Lapper is a man and Affina is a woman and answered four explicit questions that tested this gender knowledge.

Before the facts. Participants (N = 659) reported explicit beliefs that favored Lapper, the man, to be the doctor and Affina, the woman, to be the nurse \([M = -0.14, \text{one-sample } t(658) = -5.88, p < .0001, \text{Cohen’s } d = 0.23]\). Implicit beliefs likewise favored Lapper to be the doctor and Affina to be the nurse, although to a much greater extent \([M = -0.33, \text{one-sample } t(658) = -21.83, p < .0001, \text{Cohen’s } d = 0.85]\). As before, visual inspection of each belief distribution shows that explicit beliefs largely aligned with the fairness principle whereas implicit beliefs largely aligned with the base rate principle (Fig. S7).

After the facts. Participants learned individuating facts that were either stereotypic (Lapper, the man, is the doctor; Affina, the woman, is the nurse) or counterstereotypic (Affina, the woman, is the doctor; Lapper, the man, is the nurse). Once again, explicit beliefs reflected the
facts (Fig. S8) regardless of whether the facts were stereotypic \[ M_{after} = -2.62 \text{ vs. } M_{before} = -0.07; \ p < .0001 \] or counterstereotypic \[ M_{after} = 2.50 \text{ vs. } M_{before} = -0.20; \ p < .0001 \].

Implicit beliefs about Lapper and Affina’s professions, by contrast, aligned only with stereotypic facts, thereby replicating the previous results, but with novel names. After learning Affina is the doctor and Lapper is the nurse, implicit beliefs still reflected the base rate that initially favored Lapper to be the doctor (Fig. S9).

Stereotypic facts had an effect above and beyond test-retest in the expected direction \[ t(656) = -2.38, \ p = .018 \], leading to \( D \) scores consistent with Lapper’s actual profession as the doctor and Affina’s actual profession as the nurse \[ M_{after} = -0.33, \ 95\% \ CI (-0.38, -0.28) \]. While counterstereotypic facts also had an effect above and beyond test-retest \[ t(656) = 2.03, \ p = .04 \], \( D \) scores in this condition failed to reflect the fact that Affina is the doctor, not Lapper \[ M_{after} = -0.16, \ 95\% \ CI (-0.21, -0.11) \]. Lapper, the nurse, was more associated with doctor than Affina, the actual doctor.

The same pattern of results emerge when analyzing the subsample of participants \( n = 143 \) whose \( D \) scores on the first IAT were in the neutral range between -0.15 to 0.15 (Fig. S10). Among these participants who learned stereotypic facts, implicit beliefs reflected Lapper and Affina’s true professions \[ M_{after} = -0.27, \ 95\% \ CI (-0.34, -0.19) \], though the effect of these facts did not exceed test-retest \[ t(140) = -1.33, \ p = .19 \]. Among these participants who learned counterstereotypic facts, implicit beliefs still were not on the correct side of zero \( M_{after} = 0.01, \ 95\% \ CI [-0.07, 0.08] \), even though the counterstereotypic facts had an effect above and beyond test-retest \[ t(140) = 2.02, \ p = .046 \]. So as before, implicit beliefs aligned with stereotypic but not counterstereotypic facts, even among initially neutral participants.
Experiment 4

Before individuating facts are learned, explicit and implicit beliefs privilege different principles. After individuating facts are learned, explicit and implicit beliefs further dissociate. Given the ease of the explicit task, it is hardly surprising that updated explicit beliefs reflect stereotypic and counterstereotypic facts. However, it is surprising that implicit beliefs readily incorporated stereotypic facts, but not counterstereotypic facts.

In these experiments, we have assumed that the cause of the explicit-implicit dissociations is the presence of a stereotype, which does not influence explicit beliefs about specific individuals, but does influence implicit beliefs about these individuals both before and after individuating facts are learned.

In a final experiment, we directly test this assumption by using the names Matthew and Benjamin and the professions scientist and artist in the same procedure. When both individuals are male and both professions are male dominant, there is no stereotype. So explicit and implicit beliefs before individuating facts are learned should be neutral. And especially critical, once individuating facts are learned, both explicit and implicit beliefs should reflect these facts – regardless of who is actually the scientist or artist.

**Before the facts.** Before learning Matthew and Benjamin’s actual professions, participants \((N = 1417)\) reported explicit beliefs that slightly favored Matthew to be the scientist and Benjamin to be the artist \([M = -0.03, \text{one-sample } t(1416) = -2.13, p = .03, \text{Cohen’s } d = 0.06]\). Implicit beliefs, to a small degree, favored Benjamin to be the scientist and Matthew to be the artist \([M = 0.03, \text{one-sample } t(1416) = 3.03, p = .003, \text{Cohen’s } d = 0.08]\).
The large sample size magnifies this baseline difference from zero, which is negligible (Cohen’s \( d < 0.08 \)). The distributions of explicit and implicit beliefs show that the modal response is at the midpoint of zero, with the remaining responses distributed evenly on each side (Fig. S11). Thus, in the absence of a stereotype, participants displayed neutral initial beliefs.

**After the facts.** As expected, explicit beliefs (Fig. S12) were correctly updated regardless of whether Matthew turned out to be the scientist and Benjamin the artist [\( M_{\text{after}} = -2.50 \) vs. \( M_{\text{before}} = -0.05; p < .0001 \)] or vice versa [\( M_{\text{after}} = 2.16 \) vs. \( M_{\text{before}} = 0.03; p < .0001 \)].

Notably in this experiment, but not in the previous three, implicit beliefs reflected the same individuating facts as explicit beliefs did (Fig. S13). When Matthew was the scientist and Benjamin was the artist, implicit beliefs aligned with these facts [\( M_{\text{after}} = -0.06, 95\% \text{ CI} (-0.10, -0.03) \)]. The effect of these individuating facts exceeded test-retest [\( t(1414) = -4.67, p < .0001 \)]. And when Benjamin was the scientist and Matthew was the artist, implicit beliefs aligned with these facts [\( M_{\text{after}} = 0.10, 95\% \text{ CI} (0.06, 0.14) \)]. The effect of these individuating facts came extremely close to significantly exceeding test-retest [\( t(1414) = 1.95, p = 0.052 \)].

Examining the implicit beliefs of the entire sample reveals a close correspondence with the correctly updated explicit beliefs. We can see an even closer correspondence by examining only the subsample of participants (\( n = 403 \)) whose \( D \) scores on the first IAT were between -0.15 and 0.15 (Fig. 3). In this subsample, implicit beliefs reflected both types of individuating facts. When Matthew was the scientist and Benjamin was the artist, implicit beliefs aligned with these facts [\( M_{\text{after}} = -0.10, 95\% \text{ CI} (-0.14, -0.05) \)], an effect that exceeded test-retest [\( t(400) = -2.61, p = .009 \)]. And when Benjamin was the scientist and Matthew was the artist, implicit
beliefs aligned with these facts \(M_{after} = 0.09, 95\% CI (0.05, 0.14]\), an effect that also exceeded test-retest \(t(400) = 2.14, p = 0.03\). These results demonstrate that when a stereotypic base rate is absent, explicit and implicit beliefs fully converge both before and after individuating facts are learned.

**Fig. 3** Experiment 4: Mean implicit beliefs about Matthew and Benjamin’s professions among only participants whose implicit beliefs before learning the facts were neutral \((n = 403)\). IAT \(D\) scores are on the y-axis. Error bars are 95\% CIs.
Discussion

We elicited beliefs about specific individuals before and after individuating facts were learned. Before the facts were learned, explicit and implicit beliefs relied on different principles to assign each individual to a particular profession. Whereas explicit beliefs privileged the fairness principle, implicit beliefs showed excellent sensitivity to the base rate principle. After the facts were learned, explicit beliefs were amenable to individuating facts, while implicit beliefs continued to hew with base rates that counterstereotypic facts had rendered inapplicable. Stereotypes likely contributed to these differential outcomes. When both individuals were male and both professions were male dominant, the differential effects were obliterated, leading explicit and implicit beliefs to fully converge not just with each other, but also with the facts.

This research makes two main contributions. First, while past work has focused on dissociations between explicit and implicit responses, we assess how both responses compare against two kinds of real-world facts: (a) statistical regularities that can be applied and (b) individuating facts that should be applied. It is surprising that when participants only knew an individual's gender, this highly diagnostic information was largely ignored in explicitly assigning the individual to a profession. The majority of participants used this explicit question to buck stereotypes, even though it came at the cost of robust and well-known statistical likelihoods. As cultural values have shifted and will likely continue to shift, it will be important to track responses to these kinds of questions to see whether and under what conditions the base rate principle or fairness principle is deemed best to use when no individuating facts are available.
Second, this work may serve as a bridge between implicit social cognition research and models of Bayesian updating. For reasons discussed in the introduction, we forgo a formal Bayesian analysis, so we cannot make any claims about whether participants updated their beliefs in a Bayesian manner. Loosely speaking though, it seems as if explicit and implicit beliefs both conform to and deviate from Bayesian norms. For priors to be accurate, they need to reflect relevant base rates, which implicit beliefs did a far better job of than explicit beliefs did before individuating facts were learned. For posteriors to be accurate, new data need to be properly integrated with the priors. The data we presented were designed to completely overwhelm the priors, leaving no uncertainty whatsoever. When the facts were counterstereotypic, explicit beliefs gave way to the facts, but implicit beliefs did not.

Measures of explicit beliefs do little to bolster the Bayesian position, which is supported by remarkable fits between human judgment and Bayesian reasoning across a variety of domains (Griffiths & Tenenbaum, 2006, Xu & Tenenbaum, 2007). It is hardly surprising that given a clear fact about who is the doctor or scientist, responses that we have full conscious control over will properly update. However, the implicit beliefs we observed may be of interest to the Bayesian position. How can implicit beliefs shift so readily in response to data indicating a male doctor and male scientist, but not in response to data indicating a female doctor or female scientist? If the measures were geared towards judgments of entire groups, then a single counterstereotypic example need not lead to a dramatic shift in belief. But when the query is about that single example, there can only be one correct belief.

Given these findings, it may be fruitful for future research on Bayesian models of cognition to (a) work in domains that are deeply social in nature where the base rate principle...
and fairness principle can be in conflict, and (b) employ implicit measures alongside standard explicit measures to test the boundary conditions of people’s ability to update their beliefs. Much of cognition occurs unconsciously (Hassin, Uleman, & Bargh, 2005) and the topic of changing implicit responses has gained traction (Gregg et al., 2006; Reydell et al., 2007; Blair, 2002; Blair et al., 2001; Wyer, 2016; Wyer 2010; Cone & Ferguson, 2015; Lai et al., 2014). As such, it will be crucial to understand the reaches of how these implicit responses might change, and Bayesian models may be highly useful.

Two additional features make the results particularly noteworthy. First, we set a low bar for updated implicit beliefs: at minimum, $D$ scores on the second IAT should have been on the correct side of a neutral score of zero. Despite this low bar, participants in the counterstereotypic conditions failed to meet it, even when we examined those participants whose initial implicit beliefs were already on its cusp.

Second, correct implicit beliefs in the counterstereotypic conditions were absent despite the learning of highly diagnostic individuating facts. At first glance, this may appear inconsistent with Cone & Ferguson’s (2015) finding that highly diagnostic information can lead to substantial revisions of implicit evaluations. But these researchers sought to update implicit evaluations of non-stereotyped individuals, whereas we sought to update implicit beliefs about stereotyped individuals. Moreover, Cone & Ferguson found an asymmetry such that highly diagnostic negative facts were more influential than highly diagnostic positive facts. We also found an asymmetry between stereotypic vs. counterstereotypic individuating facts. Both findings dovetail with previous work demonstrating that good news vs. bad news about the self are differentially integrated into updated beliefs (Sharot et al., 2012; Eil & Rao, 2011). And taken
together with this collection of studies, the results here begin to point tentatively to boundary conditions of when implicit associations may be changed. A clearer picture of these conditions awaits future research.

The results here are also consistent with those of Reuben, Sapienza, & Zingales (2014), who found that implicit stereotypes, as measured by the IAT, predict an initial gender bias in hiring more men than women for a math task as well as a subsequent failure to correct this bias when data indicate there actually is no gender difference. This is one reason why it matters if implicit beliefs are not adequately updated to reflect the true state of the world.

But another reason is that they reveal a wide gulf between the fairness that is explicitly espoused and the ignorance that is implicitly displayed. It is humbling that the very same participants explicitly disavowed a relevant base rate and then implicitly clung to it despite clear-cut facts that had rendered it inapplicable. Insofar as this dynamic proves to be robust, this indeed is a feature of human judgment that social policies aimed at minimizing the undesired impact of stereotypes will need to take into account.

Although we have consistently demonstrated this feature of human judgment across multiple studies with large samples of participants, future research can provide additional critical tests. One issue is that the IAT may measure implicit beliefs about gendered names instead of about specific individuals. Experiment 3 – which used unfamiliar names and replicated the findings of Experiments 1 and 2 – provides strong evidence against this alternative hypothesis. However, an experiment using faces can provide further disambiguating evidence.
Unlike a novel name, which can be applied to more than one individual, a face is unique to one individual. The reason faces were not used here is that faces convey a great deal about traits – such as dominance and competence – through variations in distance between the eyes and squareness of the jaw (Todorov et al., 2015). In the present studies where professions are not only gendered but also differ in status and power and in level of skill, training, and expertise required (e.g., doctor vs. nurse), dominance and competence inferred through faces would be issues with which to contend. However, future research can and should control for these aspects and make use of faces to provide convergent or divergent evidence for the effects shown here.

Finally, consider a riddle about a father and his son who get into a car accident. The father dies on the scene and the son, who is critically injured, is transported to a hospital where the operating surgeon looks at him and exclaims, “I can’t operate on this boy – he’s my son!” In 1985, one of the authors of the present paper attempted to solve this riddle by weakly offering that perhaps the surgeon was the biological father and the other man was the adoptive father. Much to this author’s chagrin, the correct answer is that the surgeon is the boy’s mother.

Participants in Experiment 1 had no trouble giving Elizabeth an a priori equal chance to be the doctor. And when counterstereotypic facts made clear that she is actually the doctor, there was no delay in aligning explicit beliefs with the facts. And yet implicit beliefs, like the experience of puzzlement in the riddle, still indicated that Jonathan was the doctor. This association is statistically likely and important to have on hand to use as appropriate – but not when the woman turns out to be the doctor.
Materials and Methods

Institutional Approval and Informed Consent. Harvard University’s Institutional Review Board approved the experiments in this manuscript. At the beginning of each experiment, participants read and agreed to a consent form.

Participants. All participants were volunteer visitors to Project Implicit (implicit.harvard.edu). See SI for demographic information and exclusionary criteria.

Experiment 1. Participants read generic information about Jonathan and Elizabeth that revealed only their genders and that one individual is a doctor and the other is a nurse. Next, participants indicated their explicit beliefs about each individual’s profession on a Likert-type scale (-3 = Jonathan is definitely the doctor, 0 = Both individuals are equally likely to be the doctor or nurse, 3 = Elizabeth is definitely the doctor) and their implicit beliefs on an IAT in which the concepts were Jonathan, Elizabeth, Doctor, and Nurse. Participants were then randomly assigned to learn either control, stereotypic, or counterstereotypic facts about Jonathan and Elizabeth’s professions. Finally, participants indicated their explicit and implicit beliefs again on the same measures. See SI for all experimental stimuli.

IAT Scoring Procedure. Following Greenwald, Nosek, & Banaji (2003), we calculated two IAT D scores for each participant, one indicating an implicit belief before the facts were learned and a second indicating an implicit belief after the facts were learned. D scores were calculated such that negative values indicate a belief that Jonathan is the doctor (and Elizabeth is the nurse) whereas positive values indicate a belief that Elizabeth is the doctor (and Jonathan is the nurse). A D score of zero indicates a belief that both individuals are equally likely to be the doctor or nurse.
**Analyses.** All analyses were conducted using R statistical computing’s nlme package with maximum-likelihood estimation (Pinheiro et al., 2016). For both explicit and implicit beliefs, we included the interaction between time of measurement (before vs. after) and individuating facts (control vs. stereotypic vs. counterstereotypic) as a fixed effect and time of measurement nested within participant as a random effect. No other variables were included.

**Experiment 2, 3, & 4.** These experiments were identical to Experiment 1 except the names and professions were changed accordingly. In Experiment 2, participants answered four questions that tested knowledge of Lapper and Affina’s gender.

**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.
Introduction

\textit{O bir doktor.}  
\textit{O bir hemsire.}  

The Turkish pronoun \textit{o} is gender-neutral, so these sentences translate to: \textit{One is a doctor. One is a nurse.} However, Google’s (2018) translation is: \textit{He is a doctor. She is a nurse.} While statistically accurate since 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 et al., 2017), but also for anyone who makes social judgments while simultaneously attuned to group-based statistics (Garnham et al., 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:

\begin{itemize}
  \item \textit{A man performed surgery.}
  \item \textit{A woman performed surgery.}
  \item \textit{Who is more likely to be a doctor?}
\end{itemize}

The Bayesian answer is that the man is more likely to be a doctor because 1) doctors are more likely to be men, and 2) 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’ also includes procedures like 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.

\[
\frac{P(\text{Target} = \text{Doctor} \mid \text{Performed Surgery})}{P(\text{Target} \neq \text{Doctor} \mid \text{Performed Surgery})} = \frac{P(\text{Performed Surgery} \mid \text{Target} = \text{Doctor})}{P(\text{Performed Surgery} \mid \text{Target} \neq \text{Doctor})} \times \frac{P(\text{Target} = \text{Doctor})}{P(\text{Target} \neq \text{Doctor})}
\]

The prior, the far right term, is greater when the target is a man 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 far left 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 to be a doctor than the woman, conditional on both individuals performing surgery, because the man is more prototypical of the profession doctor. 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 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 (Rawls, 2001; Dworkin, 2002). Because the man and woman both performed surgery, both should be viewed
equally as doctors. This egalitarian ideal is embedded in over 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 non-profits. However, many would agree that this aspiration is not always realized. Biernat & 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 (Foschi, 1996; Rudman & Glick, 2001; Eagly & Karau, 2002). This double standard has also been implicated in job applications (Moss-Racusin et al., 2012), pay disparities (Auspurg et al., 2017), and allegations of discrimination (Price Waterhouse v. Hopkins, 1989).

Judging that the man is more likely to be a doctor than the woman when both individuals performed surgery can smack of yet another unfair double standard.

Under some circumstances, such as the Markov condition when 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 vs. 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 (McCauliffe et al., 2016) 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 assesses 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 where real money is at a stake. Furthermore, the current work establishes boundary conditions by examining a wide range of professions.

Further extending previous work, the current research elicits probability judgments and compares those judgments to what they should be according to Bayes’ 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 if 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 et al., 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 to how people evaluate someone else who makes the Bayesian judgment?

---

6 The term judgment is used refer to how likely the man vs. woman is be a doctor, conditional on both having performed surgery. Participants also made judgments of someone else who offered the Bayesian judgment; these data are discussed using the term evaluation.
**Experiment 1** assessed how people evaluate a third party that offered the Bayesian judgment in the male dominated profession of doctor.

*Procedure.* One hundred ninety nine participants (\(M_{\text{age}} = 35.62\) years, \(SD = 12.16\); 95 males, 104 females) were recruited from Amazon Mechanical Turk 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.

After learning that a man had performed surgery and that a woman had performed surgery, 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 to 7 (e.g., 1 = Extremely unfair ... 7 = 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 (see Supplemental Materials for stimuli for all studies).
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 (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 1 to 7 scales (Fig. 1B), Cronbach’s $\alpha = 0.92$, $M_{\text{composite}} = 2.28, SE = 0.10$, one-sample $t(198) = -17.24, P < 0.0001$, Cohen’s $d = 1.22$, 95% CI = [0.99, 1.53]. Even though Person X’s judgment was statistically accurate, this third party was negatively evaluated.
Experiment 2 assessed how people evaluate a third party that offered the Bayesian judgment in a wide range of male-dominated professions aside from doctor.

Procedure. Six hundred four participants (\(M_{\text{age}} = 34.56\) years, \(SD = 10.68\); 222 males, 382 females) were recruited from Amazon Mechanical Turk 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.

Participants were 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 to 7 (e.g., 1 = Extremely unfair ... 7 = Extremely fair). Last, participants provided open-ended text responses of their impressions of Person X and this person’s statement. Throughout the procedure, the order in which the man and woman were compared was randomly assigned, as was the case in Study 1.
Results. Agreement with the Bayesian judgment that the man is more likely to be a butcher, firefighter, or construction worker was at least 33% (Fig. 1A). Furthermore, evaluations of Person X along the four dimensions of fair, just, accurate, and intelligent were neutral or slightly negative (Fig. 1B). Given the high reliability of these four items within each condition, Cronbach’s αs > 0.90, all four items were averaged in each condition (see Table S1 in Supplemental Materials for item means and standard errors). These average evaluations were just below the midpoint of 4, Ms > 3.48, SEs < 0.12, and did not significantly differ from one another, Tukey adjusted Ps > 0.08, rs < 0.08, 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 U.S. are men, the proportion of construction workers who are men is greater. However, base rates alone cannot account for these results, since the gender distribution among butchers is comparable to 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 STEM professions like 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.
**Fig. 1.** Experiments 1 and 2. **A.** Proportion of participants who agreed with each judgment. **B.** Evaluations of Person X, computed by averaging the items *fair, just, accurate,* and *intelligent,* which were measured on Likert-type scales ranging from 1 to 7. Violin plots display the distributions. Error bars are 95% CIs.
Experiment 3 stringently tested if 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, Study 3 used a vignette where a) only one person, either the man or the woman, but not both, could be the doctor, b) whoever is not the doctor is 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.

Procedure. Four hundred twenty five participants were recruited from Amazon Mechanical Turk. Each participant was compensated $0.21 and could have earned up to $0.30 more. Twenty-three participants were excluded for not completing the procedure. The final sample consisted of 402 participants ($M_{age} = 34.24$ years, $SD = 10.74$; 153 males, 248 females, 1 unspecified).

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 vs. 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, based on 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 to 7 (e.g., 1 = Extremely unfair ... 7 = 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 will 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 they chose not transfer, and two 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 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, this third party was again viewed as unfair, unjust, inaccurate, and unintelligent (see Table S2 for item means and SEs) when the Bayesian
judgment was offered, as indicated by means below the midpoint of 4 on the 1-7 scales,
Cronbach’s $\alpha = 0.93$, $M_{\text{composite}} = 2.49$, $SE = 0.01$, one-sample $t(201) = -14.78$, $P < 0.0001$,
Cohen’s $d = 1.04$, 95% CI = [0.83, 1.29]. This effect was reversed (Fig. S1) 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 1-7 scales, Cronbach’s $\alpha =
0.91$, $M_{\text{composite}} = 6.34$, $SD = 0.08$, 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 to when the egalitarian judgment was offered, $M = $0.10, $SE = $0.01, $b = -$
$0.06$, $t(379.91) = 8.41$, $P < 0.0001$, Cohen’s $d = 0.84$, 95% CI = [0.63, 1.04]. Also telling are the
distributions of transfer amounts (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 proportions were reversed: 17% transferred at
least half their endowment and 65% transferred nothing.

In the Supplemental Materials, this behavioral result with real money is conceptually
replicated: participants also willingly incur a financial cost on themselves to punish someone
who makes the Bayesian judgment rather than the egalitarian judgment (Table S3 and Figs. S2-
S3). These findings lend further credence to the behavioral implications of negative evaluations.

Another study in the Supplemental Materials shows that these effects are not simply
due to the phrase “more likely” or “less likely,” which may imply a large gap. The results
replicate when these phrases are 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 Materials conceptually replicates these results in a different profession, as participants also find fault with Person X for making the Bayesian judgment that a man who communicated with air traffic control is more likely to be a pilot than a woman who communicated with air traffic control during a flight (Table S4 and Fig. S4).

Fig. 2. Experiment 3. Average amounts transferred to Person X. Errors bars are 95% CIs. Violin plots display the distribution of amounts transferred in each condition.
Experiment 4 assessed whether probability judgments would be Bayesian or egalitarian. Upon learning that a man vs. woman performed surgery, will participants judge that the man is more likely to be the doctor than the woman, or will participants judge that they are equally likely to be the doctor?

Procedure. Eight hundred ninety nine participants were recruited from Amazon Mechanical Turk 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.

Three participants were excluded because they provided priors of either 0% or 100%, which cannot be updated according to Bayes’ rule. Two participants indicated they had looked up answers to some of the questions in the study, but these participants are retained in the analyses (conclusions do not change based on whether these participants are included or excluded). While 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 ($M_{age} = 34.33$ years, $SD = 10.97$; 528 males, 364 females, 4 unspecified). Study 2 proceeded in three parts, each corresponding to a component of Bayes’ rule.
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 vs. nurse is unknown. Participants estimated the percentage chance that each person is the doctor. Since there are two hypotheses – either the man or woman is the doctor (and the other is the nurse) – both estimates had to sum to one. Thus, each participant provided his or her subjective prior about each person’s profession (e.g., the man has 75% chance to be the doctor; the woman has 25% chance to be the doctor).

2. *Posteriors.* After providing their priors, participants were randomly assigned to learn one of the following six pieces of data.

   i. The man performed surgery on a patient.

   ii. The woman performed surgery on a patient.

   iii. The man gave a sponge bath to a patient.

   iv. The woman gave a sponge bath to a patient.

   v. The man performed CPR on a patient.

   vi. The woman performed CPR on a patient.

   After learning this datum, participants again estimated the percentage chance that each person is 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 non-diagnostic of profession, as both doctors and nurses administer this procedure. For the primary analysis, only the surgery conditions (i and ii) are presented. Data from the other four conditions (iii – vi) are presented in the Supplemental Materials (Figs. S5-S6).

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. Participants were randomly assigned to estimate the corresponding likelihoods either before or after providing their subjective priors and posteriors.
Each participant’s priors and likelihoods were entered into Bayes’ rule to compute a *model posterior*, which represents what the participant’s posterior should be from a statistical perspective. This model posterior was compared against the posterior 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, $M_{\text{Man}} = 68.7\%$ vs. $M_{\text{Woman}} = 29.6\%; b = 0.39, t(890) = 17.23, P < 0.0001, r = 0.50, 95\% CI = [0.44, 0.52]$, 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, $\text{Median}_{\text{Man}} = 1.98$ vs. $\text{Median}_{\text{Woman}} = 2.65$, Wilcoxon $P = 0.19, r = 0.08$, 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 (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, $M_{\text{Model Posterior, Man}} = 87.7\%$ vs. $M_{\text{Model Posterior, Woman}} = 72.2\%$, $b = 0.15$, $t(890) = 6.83$, $P < 0.0001$, $r = 0.22$, 95% CI = [0.15, 0.27]. This disparity was also observed among participants’ reported posteriors, $M_{\text{Reported Posterior, Man}} = 86.4\%$ vs. $M_{\text{Reported Posterior, Woman}} = 78.0\%$; $b = 0.08$, $t(890) = 3.74$, $P = 0.0002$, $r = 0.12$, 95% CI = [0.05, 0.18].

In fact, relatively small differences were observed between model and reported posteriors among participants who learned that the man had performed surgery, $M_{\text{Model Posterior, Man}} = 87.7\%$ vs. $M_{\text{Reported Posterior, Man}} = 86.4\%$, $b = 0.01$, $t(1780) = 0.69$, $P = 0.49$, $r = 0.02$, 95% CI = [0.001, 0.05], and among participants who learned that the woman had performed surgery, $M_{\text{Model Posterior, Woman}} = 72.2\%$ vs. $M_{\text{Reported Posterior, Woman}} = 78.0\%$, $b = -0.06$, $t(1780) = -3.05$, $P = 0.002$, $r = 0.07$, 95% CI = [0.006, 0.16]. Thus, the posteriors reported by participants were close to the posteriors they should have reported according to Bayesian rationality (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 Materials 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 (Figs. S7-S8).
An additional study in the Supplemental Materials further demonstrates that this effect generalizes to the profession of pilots, as participants judge that a man who communicated with air traffic control during a flight is more likely to a pilot than a woman who communicated with air traffic control during a flight (Figs. S9-S13). Together, these results indicate that participants’ judgments reflect the statistical savvy of a Bayesian.
Fig. 3. Experiment 4: surgery conditions. A. Distribution of likelihood ratios (log scaled) in each condition. B. Average judgments among participants in each condition. Priors indicate judgments before participants learned that the target had performed surgery. Model posteriors indicate judgments participants should make from a Bayesian perspective. Reported posteriors indicate judgments participants actually made. Error bars are 95% CIs.
**Experiment 5** was a within-subjects design in which the same participants evaluated Person X and made probability 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.

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

The study consisted of three parts. In the first part, participants were randomly assigned to learn that either a man or 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 to 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 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 (Fig. S14). 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, $M_{\text{Model Posterior, Man}} = 94.0\%$ vs. $M_{\text{Model Posterior, Woman}} = 63.1\%$, $b = 0.31$, $t(346) = 13.53$, $P < 0.0001$, $r = 0.59$, 95% CI $= [0.52, 0.62]$. As before, this disparity was also observed among participants’ reported posteriors, $M_{\text{Reported Posterior, Man}} = 89.7\%$ vs. $M_{\text{Reported Posterior, Woman}} = 67.8\%$, $b = 0.22$, $t(346) = 9.56$, $P < 0.0001$, $r = 0.46$, 95% CI $= [0.37, 0.50]$.

Further replicating previous results, relatively small differences were observed between model posteriors and reported posteriors among participants who learned that the man had communicated with air traffic control, $M_{\text{Model Posterior, Man}} = 94.0\%$ vs. $M_{\text{Reported Posterior, Man}} = 89.7\%$, $b = 0.04$, $t(692) = 2.43$, $P = 0.02$, $r = 0.09$, 95% CI $= [0.05, 0.15]$, and among participants who had learned that the woman had communicated with air traffic control, $M_{\text{Model Posterior, Woman}} = 63.1\%$ vs. $M_{\text{Reported Posterior, Woman}} = 67.8\%$, $b = -0.05$, $t(692) = -2.64$, $P = 0.008$, $r = 0.10$, 95% CI $= [0.01, 0.21]$. 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 identically problem, albeit the effect was weakened since base rates
were made salient by the first parts of the procedure. Seventy nine percent of participants agreed that a man and woman are equally likely to be a doctor given that they both performed surgery, 20% agreed that the man is more likely to be a doctor, and 1% agreed that the woman is more likely to be a doctor. In the Supplemental Materials, additional analyses reveal the proportion of participants (71%) who used gendered base rates when making their probability judgments but not when indicating which judgment they agreed with (Table S5).

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 1-7 scales, Cronbach’s \( \alpha = 0.91 \), \( M_{\text{composite}} = 3.29, SE = 0.08 \), one-sample \( t(347) = -8.90, P < 0.0001 \), 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 vs. woman is the pilot conditional on having communicated with air traffic control (ATC). Perhaps participants who criticized Person X also made egalitarian judgments that give the man the woman equal probabilities of being the pilot.

But as shown in Fig. 4, participants judged that the man is more likely to be the pilot than the woman, regardless of their evaluation of Person X, \( F(1, 344) = 84.58, P < 0.0001, \eta^2 = 0.19 \), 95% CI = [0.13, 0.27]. 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 is more likely to be the
pilot than the woman, \( Fitted_{\text{Reported Posterior, Man}} = 91.6\% \) vs. \( Fitted_{\text{Reported Posterior, Woman}} = 81.4\% \), \( b = -0.10, SE = 0.04, t(344) = -2.27, P = 0.02, r = 0.12 \) 95% CI = [0.03, 0.21].

The difference in probability judgments of the male and female targets increases as evaluations of Person X become more positive, \( F(1, 344) = 10.71, P = 0.001, \eta^2 = 0.02, 95\% CI = [0.005, 0.07] \). However, participants were equally and highly accurate irrespective of how they felt towards Person X, as evidenced by the minimal difference between their model and reported posteriors across the entire range of evaluations (Fig. S15). Thus, participants accurately judged that the man is more likely to be the pilot than the woman; these participants then criticized Person X for making a conceptually similar Bayesian judgment. A study in the Supplemental Materials replicates this key analysis when Bayesian judgments are elicited through the doctor scenario and Person X’s statement concerned who is more likely to be the pilot (Figs. S16-S18).
Fig. 4. Experiment 5. Reported posterior probabilities as a function of evaluations of Person X (average of four Likert-type items). Grey bands are $SE$s.
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 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 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 Tetlock et al. (2000), who coined the term forbidden base rates to refer to statistics that some many 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. An unabashed sexist and a feminist statistician can both state that a man who performed surgery is more likely to be a doctor than a woman who performed
surgery, albeit for different reasons. Given this uncertainty, participants may have attended to
the base rate that Person X is \textit{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 vs. 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 \textit{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 upon efforts to formalize the process by which the
motivation to assess character relates to probability judgments (Pizarro & Tenenbaum, 2012;
Kleiman-Weiner et al., 2017).

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 to be the doctor than the woman. It may be the case, then, that the same cognitive
process of Bayesian inference underlies statistical judgments and negative evaluations of others
who make certain statistical judgments.

One limitation here is that data were collected from Amazon Mechanical Turk, 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 Mechanical Turk are comparable to results obtained from laboratory settings (Amir, Rand, & Gal, 2012).

Finally, these findings have implications for criminal trials where 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. While 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 providing the equal treatment they desire.
PAPER 3


Abstract

From a statistical standpoint, judgments about an individual are more accurate if base rates about the individual’s social group are taken into account (Eddy, 1982; Kahneman & Tversky, 1973; Bar-Hillel, 1980; Tversky & Kahneman, 1974). But from a moral standpoint, using these base rates is considered unfair and can even be illegal (Cao & Banaji, 2016; Rawls, 2001; Dworkin, 2000; Koehler, 1992; Test-Achats v. Council of Ministers, 2011). Thus, the imperative to be *statistically accurate* is directly at odds with the imperative to be *morally fair*. This conflict was resolved by creating tasks in which Bayesian rationality and moral fairness were aligned, thereby allowing social judgments to be both accurate and fair. Despite this alignment, we show that social judgments were *inaccurate* and *unfair*. Instead of appropriately setting aside social group differences, participants erroneously relied upon them when making judgments about specific individuals. This bias – which we call *base rate intrusion* – was robust, generalized across various social groups (gender, race, nationality, and age), and differed from analogous nonsocial judgments. Results also demonstrate how social judgments can be corrected to achieve both statistical accuracy and moral fairness. Overall, these data (total *N* = 5,138) highlight the pernicious effects of social base rates: under conditions that closely approximate those of everyday life (Fiske & Neuberg, 2012; Moss-Racusin et al., 2012; Cheryan, et al., 2009), these base rates can undermine the rationality and fairness of human judgments.
Introduction

Many studies illustrate the importance of base rates for making accurate judgments. To assess how likely a woman is to have breast cancer given that her mammogram results are positive, the prevalence of the disease must be considered (Eddy, 1982). To determine if a man who enjoys mathematical puzzles is an engineer or lawyer, the distribution of these professions among the man’s group is relevant (Kahneman & Tversky, 1973). However, base rates are often inadequately weighed or outright ignored; this error is called base rate neglect and has been shown to undermine the accuracy of human judgments (Bar-Hillel, 1980; Tversky & Kahneman, 1974).

The current work focuses on a specific type of base rate: stereotypes about social groups. Like any base rate, using stereotypes can increase the probability that a judgment about an individual will be accurate, as evidenced by decades of research conceptualizing stereotypes as base rates (Locksley et al., 1980; Rasinski et al., 1985; Hamilton, 1981; Krosnick et al., 1990; Jussim, 2012). But unlike other base rates, using stereotypes raises serious questions about fairness (Cao & Banaji, 2016). Many theories of morality eschew the application of group characteristics to specific individuals because doing so violates individual rights and basic tenets of justice (Rawls, 2001; Dworkin, 2002). In fact, Western democracies have codified this position. For instance, despite the diagnosticity of base rates that emanate from group membership, they cannot be used to decide guilt in American courtrooms (Koehler, 1992), nor can they be used to determine insurance premiums in the European Union (Test-Achats v. Council of Ministers, 2011). Thus, a clear tension emerges between two imperatives.
To uphold statistical accuracy can be perceived as undermining moral fairness. But to uphold moral fairness can be perceived as committing the blunder of base rate neglect.

Here, we completely remove this tension between accuracy and fairness by creating tasks in which Bayesian rationality dictates that base rates should *not* be used. When base rates are rendered statistically irrelevant, base rate neglect is no longer an error but a dual prescription: base rates that differentiate between two social groups should be ignored because doing so achieves both accuracy and fairness. In other words, any intrusion of base rates into judgments about another individual would be irrational from a Bayesian standpoint and unfair from a moral standpoint.

The following example illustrates how accuracy and fairness can be simultaneously achieved. Consider the base rate that doctors tend to be male, and the base rate that nurses tend to be female. Now imagine a charity that invites medical professionals to an event based *solely* on whether they are a doctor or nurse. If someone is a doctor, that person is likely to be invited. If someone is a nurse, that person is unlikely to be invited. Given these premises, the charity is more likely to invite a male than a female since the former is more likely to be a doctor. However, this is only the case when the charity does not know if the person in question is a doctor or nurse. Once the person’s profession becomes known, gender ceases to be of relevance and therefore should *not* be used. With respect to who will be invited, a female doctor should be treated the same as a male doctor – even though doctors tend to be male. Likewise, a male nurse should be treated the same as a female nurse – even though nurses tend to be female.
This reasoning is formalized in Fig. 1A as a Bayesian network, a directed acyclic graph where nodes are variables and arrows are causal influences (Pearl, 2000). This specific network structure is called a chain, and its properties dictate that once the middle node is known, the top and bottom nodes become independent, thereby rendering the top node irrelevant to judgments about the bottom node. This is an example of the Markov assumption (Hausman & Woodard, 1999), which specifies when variables become conditionally independent of one another (see Supplementary Information for probability calculus). While previous work has shown that people violate the Markov assumption (Rottman & Hastie, 2014; Rehder, 2014), the current work systematically tests base rates from the social vs. nonsocial domains.

As discussed above, social base rates present a tension between accuracy and fairness, which the Markov assumption resolves. This tension, however, does not arise when base rates concern nonsocial entities. Consider the base rate that intact spoons tend to be made from metal, and the base rate that broken spoons tend to be made from plastic. Using this base rate to judge the future outcome of a spoon has consequences for accuracy, but not for fairness since no individual rights are violated. But whether social vs. nonsocial base rates fundamentally differ in their influence on human judgments is unclear. On the one hand, both types of base rates help simplify a complex world (Murphy, 2002; Medin & Smith, 1984), which raises possibility that social and nonsocial knowledge share the same cognitive underpinnings (Hamilton, 1981; Banaji & Bhaskar, 1999). On the other hand, social knowledge is often infused with greater emotion (Norris et al., 2004) and is differentially structured compared to knowledge about physical objects (Wattenmaker, 1995). Indeed, social neuroscientists have found that thinking about social entities activates a distinct set of brain regions relative to
thinking about nonsocial entities, suggesting that different mechanisms may support these two processes (Contreras et al., 2012; Mitchell et al., 2002).

To investigate the potentially unique standing of social base rates vis-à-vis human judgments, we constructed a scenario that was exactly parallel in structure to the gender scenario, except the base rates concerned intact spoons, which tend to be made from metal, and broken spoons, which tend to be made from plastic (Fig. 1B). Imagine a factory that makes spoons and decides which ones to keep based solely on whether they are intact or broken. If the spoon is intact, it is likely to be kept. If the spoon is broken, it is unlikely to be kept. Once the factory knows that a particular spoon is intact or broken, its material (metal vs. plastic) should not influence whether it is kept. That is, the base rate should not be used. With respect to which spoon will be kept, an intact plastic spoon should be treated the same as an intact metal spoon – even though intact spoons tend to be made from metal. Likewise, a broken metal spoon should be treated the same as a broken plastic spoon – even though broken spoons tend to be made from plastic.

In the following experiments, participants (total $N = 5,138$) were randomly presented with a scenario whose logical structure is depicted in either Fig. 1A or 1B. The wording of the scenarios was adapted from Krynski & Tenenbaum (2007), who established wording that conveyed a chain network structure. After reading the scenario, each participant made two judgments about the bottom node given knowledge of the top and middle nodes. The only difference between these two judgments was what state of the top node was known. In the spoon scenario, participants judged the likelihood of keeping a broken metal spoon vs. a broken plastic spoon, or the likelihood of keeping an intact metal spoon vs. an intact plastic spoon. In
the gender scenario, participants judged the likelihood of inviting a male nurse vs. a female nurse, or the likelihood of inviting a male doctor vs. a female doctor. If base rates are properly set aside, both judgments on the 1-7 Likert-type scale ($1 = \text{Extremely unlikely} \ldots 7 = \text{Extremely likely}$) should be the same.

**Fig. 1.** Two Bayesian networks that are identical in structure but differ in whether the base rates are social vs. nonsocial. In (A), gender influences an individual’s profession, which influences whether he or she will be invited by a charity. In (B), a spoon’s material influences whether it remains intact or breaks, which influences whether it will be kept by a factory.
Experiment 1

In Experiment 1, the Markov assumption was violated, a result that replicates previous work (Rottman & Hastie, 2014; Rehder, 2014). Judgments were influenced by a spoon’s material or an individual’s gender, when in actuality, this information should have been set aside (Fig. 2). Notably, however, these violations strongly depended on whether the base rates were nonsocial or social \( F(1, 395) = 63.06, P < 0.0001 \). In the spoon scenario, which contained nonsocial base rates, the plastic spoon was judged less likely to be kept regardless of whether the two spoons in question were broken \( M_{\text{broken metal}} = 4.91 \) vs. \( M_{\text{broken plastic}} = 3.73; b = 1.19, t(395) = 5.51, P < 0.0001 \) or intact \( M_{\text{intact metal}} = 6.52 \) vs. \( M_{\text{intact plastic}} = 5.10; b = 1.42, t(395) = 6.53, P < 0.0001 \). But in the gender scenario, which contained social base rates, judgments erroneously relied upon group differences. A male nurse was judged less likely to be invited than a female nurse \( M_{\text{male nurse}} = 3.70 \) vs. \( M_{\text{female nurse}} = 5.58; b = -1.88, t(395) = -8.69, P < 0.0001 \), but a female doctor was judged less likely to be invited than a male doctor \( M_{\text{male doctor}} = 6.18 \) vs. \( M_{\text{female doctor}} = 4.36; b = 1.81, t(395) = 8.16, P < 0.0001 \). Despite having the opportunity to be both statistically accurate and morally fair, social judgments broke with Bayesian rationality and with tenets of fairness. The small main effect of profession (nurse vs. doctor) in the social condition raises the possibility that participants may not have comprehended the scenarios. We rule out this possibility by replicating Experiment 1 and including a comprehension check (Supplementary Figs. 1 and 2).
**Fig. 2.** Experiment 1 ($N = 399$). Participants’ average judgments when nonsocial vs. social base rates should not have been used. Each participant was randomly assigned to make two judgments on a 1 to 7 scale (e.g., broken metal spoon vs. broken plastic spoon). All possible pairs of judgments a participant could have been randomly assigned to make are on the x-axes. Statistical tests compare the means of each pair of judgments. Points show the distribution of judgments. Error bars are 95% CIs. *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, n.s. $P > 0.05$
Experiment 2

Experiment 2 sought to replicate the findings and ensure that they were not specific to nonsocial base rates about spoons or to social base rates about gender. These base rates were tested once more alongside other nonsocial base rates about topics as wide ranging as the weather, days of the week, and alarms. Social base rates were also richly varied and concerned race (black vs. white), nationality (American vs. Foreign), and age (young vs. old). Visual inspection indicates that the results replicate, demonstrating the generalizability of the findings (Supplementary Figs. 3 and 4). The results of the alarm scenario appear similar to the results of scenarios containing social base rates. This similarity may have emerged because this scenario blends aspects from both the social and nonsocial domains. Alarms emanate from inanimate physical objects, but burglary is an activity perpetrated by one group (burglars) against another (tenants/homeowners). This blending raises the possibility that social and nonsocial content are on a continuum, and that the alarm scenario may be located near a blurry boundary that future research may explore.

The results of Experiment 2 were collapsed according to whether base rates were nonsocial or social (Supplementary Fig. 5). When base rates were nonsocial, judgments erred such that, for example, a plastic spoon was judged less likely to be kept regardless of whether the spoons in questions were broken or intact. But when base rates were social, group differences in gender, race, nationality, or age impinged on judgments that upheld neither Bayesian rationality nor tenets of fairness – even though upholding both was possible.
Experiment 3

When two individuals differed in gender but not in profession, participants’ judgments erroneously relied upon gender differences. As a robustness test of this effect, participants in Experiment 3 made judgments about two individuals who differed in profession but not in gender. That is, participants judged how likely a male nurse vs. a male doctor were to be invited, or they judged how likely a female nurse vs. female doctor were to be invited. Without a contrast gender, perhaps participants would base their judgments on the logical structure of the task instead of on group differences between men and women. However, the same incorrect reliance on group differences was observed once again (Supplementary Fig. 6), as was the contrast between the nonsocial and social conditions \([F(1, 394) = 126.90, P < 0.0001]\). When the base rates were social, a male nurse was judged less likely to be invited than a female nurse \([M_{\text{male nurse}} = 3.55 \text{ vs. } M_{\text{female nurse}} = 5.81; b = -2.25, t(394) = -11.49, P < 0.0001]\), but a female doctor was judged less likely to be invited than a male doctor \([M_{\text{male doctor}} = 6.25 \text{ vs. } M_{\text{female doctor}} = 3.82; b = 2.43, t(394) = 12.40, P < 0.0001]\). Again, social judgments upheld neither Bayesian rationality nor tenets of fairness, even though both accuracy and fairness were achievable.
Experiment 4

Having established the generalizability and robustness of the effect, we next set out to correct it. Can social judgments adhere to the Markov assumption and therefore achieve both statistical accuracy and moral fairness? Demonstrating the effort required to accomplish both ends can illuminate the tenacity of social base rates. To create the strongest conditions that would enable social judgments to be both accurate and fair, three potential problems with the experiments thus far were identified and remedied simultaneously in Experiment 4.

I. The gender scenario stated that more males than females would be invited. Likewise, the spoon scenario stated that more metal spoons than plastic spoons would be kept. These statements about disparate outcomes may have led participants to erroneously perpetuate them. To prevent participants from committing this naturalistic fallacy (Kay et al., 2009) – confusing what is the case for what ought to occur – these statements were removed in Experiment 4.

II. The gender scenario explicitly referenced the likely medical professions of males vs. females. Likewise, the spoon scenario explicitly referenced different breakage rates between metal vs. plastic spoons. These explicit base rate references may have led participants to assume that base rates were relevant to judgments. To prevent participants from following this Gricean maxim of relevance (Grice, 1975) – assuming that all information provided is pertinent – explicit references to base rates were also removed in Experiment 4.

III. In both the gender and spoon scenarios, judgments should have been about the bottom node given knowledge of the top and middle nodes. However, participants may have
done the opposite by making judgments about the top and middle nodes given knowledge of the bottom node. To prevent participants from committing this inverse fallacy (Villejoubert & Mandel, 2002) – misinterpreting the judgment to be made as information already known – the persons and spoons in question were uniquely individuated in Experiment 4, which made abundantly clear what information was known and what judgment needed to be made.

When all three strategies were implemented simultaneously, social judgments were both accurate and fair (Fig. 3). Although nonsocial judgments were still erroneously influenced by a spoon’s material, a person’s gender no longer substantially influenced participants’ social judgments [$F(1, 407) = 60.15, P < 0.0001$]. With respect to who would be invited, parity was achieved between a male nurse and female nurse [$M_{\text{male nurse}} = 4.46$ vs. $M_{\text{female nurse}} = 4.61$; $b = -0.15$, $t(407) = -1.32$, $P = 0.19$] and between a female doctor and male doctor [$M_{\text{male doctor}} = 5.76$ vs. $M_{\text{female doctor}} = 5.51$; $b = 0.25$, $t(407) = 2.29$, $P = 0.02$]. Gender-profession base rates were kept at bay, resulting in social judgments that were consistent with Bayesian rationality and with tenets of fairness.
Fig. 3. Experiment 4 ($N = 411$). Participants’ average judgments when non-social vs. social base rates should not have been used. Each participant was randomly assigned to make two judgments on a 1 to 7 scale (e.g., broken metal spoon vs. broken plastic spoon). All possible pairs of judgments a participant could have been randomly assigned to make are on the x-axes. Statistical tests compare the means of each pair of judgments. Points show the distribution of judgments. Error bars are 95% CIs. *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, n.s. $P > 0.05$. When all three strategies were implemented, the effect remains in non-social judgments, but social judgments become accurate and fair.
Discussion

To achieve parity between two individuals of the same profession but different gender, all three strategies needed to be implemented simultaneously. When just one strategy or even pairs of strategies were implemented, social judgments improved somewhat but still erroneously relied upon group differences that should have been disregarded, showing that this finding is not merely an instantiation of other cognitive biases (see Experiments 5-7 in Supplementary Information and Supplementary Figs. 7-9). Comparing the relative efficacy of the three aforementioned strategies tentatively suggests that removing explicit references to base rates and individuating the persons in question are particularly helpful for achieving both accuracy and fairness in social judgments (Supplementary Fig. 10). Although future research is needed to confirm this finding, it is consistent with Fiske & Neuberg’s (1990) model of impression formation: removing explicit references to base rates may decrease the activation of group stereotypes, and individuating the persons in question may further decrease reliance on stereotypes. In conjunction, these strategies could reliably prevent the improper intrusion of base rates into social judgments.

The results of Experiment 4 also underscore the extensive work required to construct a representation of the task that enables social judgments to be accurate and fair. Although the tasks in Exps. 1-3 lack the remedies that together eliminated the bias, these earlier tasks faithfully represent the conditions under which social judgments are typically made. In everyday life, disparate outcomes between social groups are common (Moss-Racusin et al., 2012); base rates about these groups are exaggerated in stereotype knowledge (Cheryan et al., 2009); and stereotyped people are not always uniquely individuated (Fiske & Neuberg, 1990).
However, the presence of any one of these features can lead to the mistaken use of gender, race, nationality, or age in social judgments.

We call this phenomenon *base rate intrusion*, which can be conceptualized as the opposite of base rate neglect. Under many conditions, base rates should be entered into judgments of statistical likelihood (Eddy, 1982; Kahneman & Tversky, 1973; Bar-Hillel, 1980; Tversky & Kahneman, 1974). But when the conditional independence structure of a task requires that base rates be set aside, any intrusion of base rates would be mistaken. Base rate intrusion is especially pernicious in the social domain because it encapsulates not one but two phenomena that can be considered errors. First, Bayesian rationality is undermined. Second, tenets of fairness are violated. A male nurse and female nurse should be judged equally. Likewise, a female doctor and male doctor should also be judged equally. This parity satisfies both Bayesian rationality *and* basic tenets fairness. Despite having the opportunity to uphold both of these often-opposed normative standards, social judgments fell short on both accounts.

This bias was observed not only in participants’ average judgments, but also in contour plots that depict which 1-7 likelihood judgments were popular among the large samples of participants tested across the wide range of scenarios in Experiment 2 (Supplementary Fig. 12). Nearly all participants fell prey to base rate intrusion by incorporating irrelevant group differences into their judgments. These contour plots also refute an alternative account, namely that stereotype-incongruent targets (e.g., male nurses, female doctors) may have decreased confidence or induced confusion among participants (Rottman & Hastie, 2016). If this were the case, then there should be a high density of participants who gave judgments at the low end (1) or midpoint (4) of the scale. However, 1s or 4s were hardly observed for
judgments about stereotype-incongruent targets. Instead, the distribution of judgments is consistent with the erroneous use of base rates.

These contour plots also reveal a different error pattern for nonsocial judgments, relative to social judgments. Plastic spoons, for example, were judged less likely to be kept than metal spoons regardless of whether the spoons were broken or intact. This tendency is reflected in participants’ average judgments. However, many participants were able avoid this error, as there are clusters of participants whose judgments lie on the 45-degree identity line (Supplementary Fig. 13). These findings further suggest important differences in how social vs. nonsocial base rates influence human judgments despite the parallel logic of the tasks.

One possible reason for this differential influence is that social base rates may be richer, more familiar, or more accessible than nonsocial base rates. Consequently, the likely medical professions of men vs. women, for example, may hold considerable sway over social judgments, leading to robust judgments that favor female nurses over male nurses but male doctors over female doctors. However, strategies that virtually eliminated base rate intrusion in social judgments did not have the same ameliorating effect on nonsocial judgments (Supplementary Fig. 11). This result raises the possibility that social base rates – even if they are richer, more familiar, or more accessible – may not be as entrenched as nonsocial base rates. Knowledge of social group differences raises questions of fairness, perhaps motivating people to disregard this knowledge if the judgment task is constructed as it was in Experiment 4. Current theories of causal reasoning are agnostic to semantic content (Rottman & Hastie, 2014; Rehder, 2014), so future research is needed to test these and other possibilities, which, alongside the findings
presented here, could further refine theories of how the human mind constructs and uses Bayesian networks to reason about the social vs. nonsocial domains.

The social domain is of particular interest because it is here where the twin goals of statistical accuracy and moral fairness converge. It is not always the case that base rates are statistically relevant. So when Bayesian rationality dictates that base rates should be disregarded, an accurate judgment also becomes a fair judgment. But under conditions that closely approximate those of everyday life, we have shown that group differences are improperly used, which undermines both statistical rationality and basic fairness in judgments about other people.
Methods

Participants, sample size, and informed consent. All participants were recruited from Amazon Mechanical Turk. Large sample sizes of approximately 100 participants per cell were determined a priori based on past research (see Supplementary Information for demographic information and sample size, mean, and SD for each condition in all experiments). Harvard University’s Institutional Review Board approved the experiments in this paper. All experiments complied with relevant ethical regulations, and informed consent was obtained from all participants.

Procedure. In all experiments except Experiment 3, participants were randomly assigned to either a social or nonsocial scenario (see Supplementary Information for stimuli). After reading the scenario, each participant made two judgments about the bottom node given knowledge of the top and middle nodes (see Figs. 1A and 1B). The only difference between these two judgments was what state of the top node was known. In the nonsocial scenario about spoons, participants were randomly assigned to judge the likelihood of keeping a broken metal spoon vs. a broken plastic spoon, or the likelihood of keeping an intact metal spoon vs. an intact plastic spoon. In the social scenario about gender, participants were randomly assigned to judge the likelihood of inviting a male nurse vs. a female nurse, or the likelihood of inviting a male doctor vs. a female doctor. All judgments were made on the same 1-7 Likert-type scale (1 = Extremely unlikely … 7 = Extremely likely). To remove any possible memory effects, participants were able to refer to the scenario when making their judgments. Social and nonsocial content were varied in Experiment 2. In Experiment 3, participants again made two
judgments, but they were conditioned on the same state of the top node but different states of the middle node. Each experiment was conducted once.

**Analyses.** Analyses for all experiments were conducted using R statistical computing’s `nlme` package (Pinheiro et al., 2017). The three-way interaction between base rate (nonsocial vs. social), the middle node in the chain network (broken/nurse vs. intact/doctor), and the top node in the chain network (plastic/female vs. metal/male) was included as a fixed effect. The top node nested within participant was included as a random effect. No other variables were included. In Experiment 3, the between and within-subjects conditions were switched, so the fixed effect remained the same while the random effect was changed to the middle node nested within participant. In Experiment 2, the `lme4` package was also used (Bates et al., 2015). The fixed effect was again the three-way interaction between base rate type, the middle node, and top node. Random effects for participant and the various scenarios that were tested were also included. All statistical tests were two-sided.
GENERAL DISCUSSION

The three papers in this dissertation investigated the judgments people make about specific individuals when Bayesian principles and egalitarian values are both at stake. Consistent across each paper was the finding that people employed base rates even though they believed that it was inappropriate to do so. In Paper 1, participants eschewed obvious and relevant base rates in their explicit judgments but hewed with these base rates in their implicit judgments – both before and after individuating information was learned. In Paper 2, participants shared less money with a third party who made a judgment that was consistent with established base rates. These participants even went as far as to incur financial costs on themselves to punish this third party. However, these very same participants used base rates as a Bayesian statistician would to render the same judgment that they found morally and intellectually repugnant when offered by someone else. Finally, in Paper 3, participants continued to use base rates even when Bayesian reasoning dictated that these base rates should be ignored. This resulted in participants committing not only a statistical error, but also a moral error according to their own standards of conduct.

A common vocabulary for Bayesian principles and egalitarian values

A challenge in studying Bayesian principles and egalitarian values is that each set of concepts relies on a distinct vocabulary. Bayesian principles are captured by mathematical symbols and statistical theory, whereas egalitarian values are supported by legal doctrine and moral philosophy. There would appear to be a wide gulf separating the two sets of concepts, which can smack of comparing proverbial apples and oranges.
The reason this vocabulary gap is important is that the proliferation of artificial intelligence has raised ethical concerns and increased calls for transparency and oversight (O’Neil, 2016; Fry, 2018). Across the United States, algorithms are used to decide the bail eligibility of defendants, the loan worthiness of credit applicants, and which advertisements appear on the computer and smartphone screens of millions of individuals. Given the vast scale of these enterprises and the potential for harm against already marginalized groups, the importance of what has been termed algorithmic fairness has never been higher.

As this dissertation has shown, people hold strong convictions about what is fair and what is not. However, mathematically formalizing these convictions and translating them into lines of code that can then be compiled comprise a tall but necessary order for promoting algorithmic fairness. Machines, after all, do not speak human.

Is it possible, then, for Bayesian principles and egalitarian values to be expressed in a common language, and, if so, what might psychological science have to contribute? In short, the answers are yes and quite a bit. To begin, egalitarianism implies equality between individuals of different social groups. In the papers presented here, this equality failed to emerge, instead supplanted by unequal judgments that relied on base rates. These results came from experiments that share a common design with many experiments in social psychology, particularly those from the stereotyping and prejudice literature.

In this common design, participants learn the group membership of a target individual, which is systematically manipulated (e.g., gender). Participants also learn information specific to the individual, which is constant across conditions (e.g., performed surgery), and from an egalitarian perspective, is the only appropriate basis for the outcome that comprises the
dependent variable. This type of design can be summarized in the following conditional probability:

\[ P(\text{outcome} \mid \text{group membership, individuating information}) \]

In Paper 2, for example, a key planned contrast revealed the following:

\[ P(\text{doctor} \mid \text{female, performed surgery}) < P(\text{doctor} \mid \text{male, performed surgery}). \]

Statisticians would characterize each probability as posterior predictive value (PPV), as \textit{performed surgery} is new information that allows a prior to be updated to a posterior. The observed disparity between PPVs is one instance of egalitarianism failing to materialize. But if the PPVs had been equivalent across both conditions, then egalitarianism would have emerged. Thus, parity between PPVs is one way in which egalitarianism can be expressed in the same mathematical symbols as Bayesian principles.

Critically, these symbols can also be used to express different forms of egalitarianism. Instead of testing for an outcome conditional on group membership and new information – as in the case of PPV – an experiment can instead test for an outcome conditional on group membership and the ground truth. The current work did not employ this exact design, but previous research has. For example, in the shooter bias paradigm, participants were tasked with not shooting unarmed avatars that were either Black or White (Correll, Park, Judd, & Wittenbrink, 2007). The outcome – whether to shoot or not – was based on group membership and the ground truth that the avatars were unarmed. While an egalitarian result would have been no difference in the likelihood of shooting an unarmed Black vs. White avatar, this was
not the case, as participants were more likely to mistakenly shoot unarmed Black avatars than unarmed White avatars. This difference can be expressed as $P(\text{shoot} \mid \text{White, unarmed}) < P(\text{shoot} \mid \text{Black, unarmed})$. Here, each conditional probability is a false positive rate (FPR), and the disparity between FPRs is another instance of egalitarianism failing to materialize and expressed in the same language as Bayesian principles.

Egalitarianism as parity between FPRs naturally calls to mind the opposite form: egalitarianism as parity between false negative rates (FNRs). An example of this disparity comes from medical domain: despite suffering from comparable degrees of pain, Black patients are more likely to be undertreated than White patients (Anderson, Green, & Payne, 2009), a result that, once again, can expressed using symbols native to Bayesian principles:

$$P(\text{undertreat} \mid \text{White, in pain}) < P(\text{undertreat} \mid \text{Black, in pain}).$$

**Tradeoffs in egalitarianism**

Common throughout psychology and more broadly in the social sciences are demonstrations of people failing to uphold egalitarian values. As explained above, these failures can be sorted into one of three categories: disparities in PPVs, FPRs, or FNRs. Collectively, these findings underscore the urgency of correcting these disparities to produce PPVs, FPRs, and FNRs that are equivalent irrespective of group membership. And as this dissertation has shown, the very same people who espouse egalitarianism can be the very same ones who undermine it by relying on base rates. But a further wrinkle that will require additional research to sort out is which conception of egalitarianism – equal PPVs, FPRs, or FNRs – people value most.
Many would agree that in an ideal world, all three conceptions of egalitarianism would be fulfilled. So it might seem odd to pose the question of which conception people value most. However, all three conceptions cannot necessarily be achieved at once. At least one conception of egalitarianism may have to be left unfulfilled if base rates between social groups differ (Kleinberg, Mullainathan, & Raghavan, 2016). To illustrate this constraint, consider, for one final time, Kahneman & Tversky’s (1973) mammogram problem, whose solution is below in Bayes’ theorem (pos = positive mammogram):

\[
P(cancer|pos) = \frac{P(pos|cancer)P(cancer)}{P(pos|cancer)P(cancer) + P(pos|no\ cancer)P(no\ cancer)}
\]

First, it is useful to rewrite this equation in terms of the PPV, FPR, and FNR. Note that the term on the far left, \(P(cancer \mid pos)\), is the PPV and that \(P(cancer)\) is the base rate. The base rate will henceforth be abbreviated as \(BR\), which means that \(P(no\ cancer)\) can be written as \(1 - BR\). Furthermore, \(P(pos\mid no\ cancner)\) is the false positive rate, so it will be abbreviated as FPR. And finally, \(P(pos\mid cancer)\) can be written as its inverse, becoming \(1 - P(neg\mid cancner)\), which is equivalent to \(1 - FNR\) since \(P(neg\mid cancner)\) is the false negative rate. With this in mind, the equation can now be rewritten as the following:

\[
PPV = \frac{(1 - FNR) \times BR}{(1 - FNR) \times BR + FPR \times BR}
\]
After algebraic manipulation to solve for FPR, the same equation becomes:

\[
FPR = \frac{BR}{(1 - BR)} \times \frac{(1 - PPV)}{PPV} \times (1 - FNR)
\]

From here, we will operate on two versions of the above equation, one for women (subscript \(W\)) and another for men (subscript \(WM\) to show why there are tradeoffs in egalitarianism:

\[
FPR_W = \frac{BR_W}{(1 - BR_W)} \times \frac{(1 - PPV_W)}{PPV_W} \times (1 - FNR_W)
\]

\[
FPR_M = \frac{BR_M}{(1 - BR_M)} \times \frac{(1 - PPV_M)}{PPV_M} \times (1 - FNR_M)
\]

To assess egalitarianism in this context, we will compare false positive rates between women and men, which can be accomplished by forming a ratio:

\[
\frac{FPR_W}{FPR_M} = \frac{BR_W}{(1 - BR_W)} \times \frac{(1 - PPV_W)}{PPV_W} \times (1 - FNR_W)
\]

Setting up the equation in this manner allows us momentarily to fulfill two conceptions of egalitarianism: equal PPVs and equal FNRs for men and women. Parity in these constructs between men and women would allow these terms to cancel out, resulting in simply:
In turn, this can be written as:

\[
\frac{FPR_W}{FPR_M} = \frac{BR_W}{(1 - BR_W)} \times \frac{BR_M}{(1 - BR_M)}
\]

From the above equation, one can see that the false positive rates between men and woman can only be equal if the base rates between men and woman are also equal. Recall that the base rate refers to \( P(\text{cancer}) \), specifically breast cancer, which raises the question of whether it is realistic to assume that the prevalence of breast cancer is equal between men and women. Surely many would agree that this assumption is unrealistic, which means that equating PPVs and FNRs between men and women results in a disparity in FNRs.

This specific tradeoff, as illustrated in the left panel of Figure 1 below, worsens as the base rate difference between men and women increases. The same phenomenon occurs when the two remaining pairwise combinations of egalitarian conceptions are met: the third cannot be met due to differences in base rates, as shown in the middle and right hand panels.
When two conceptions of egalitarianism are achieved, the third cannot be achieved when base rates between two groups differ. PPV = posterior predictive value; FPR = false positive rate; FNR = false negative rate.

The implications of this tradeoff loom large. A recent report by investigative journalists revealed biases in how Black and White defendants were treated by an algorithm that determines bail eligibility (Angwin, Larson, Mattu, & Kirchner, 2016). Specifically, these biases were disparities in FPRs and FNRs: after posting bail, Black defendants who were not rearrested were more likely to be deemed “high risk” than White defendants who were not rearrested, while White defendants who were rearrested were more likely to be deemed “low risk” than Black defendants who were rearrested.

This report caused quite a stir, including a lively defense of the algorithm by criminal justice researchers showing that the PPVs were virtually indistinguishable for Black and White defendants: given a certain risk score, Black and White defendants were equally likely to be rearrested (Flores, Bechtel, & Lowenkamp, 2016). What was missing from this volley was the insight that not all three conceptions of egalitarianism could be simultaneously met. Likely because of systemic factors that result in over-policing of Black individuals, base rate differences in arrest rates have forced tradeoffs in egalitarianism.
Which conceptions of egalitarianism do people value more?

Of course, equality in PPVs, FPRs, and FNRs are not the only ways of conceptualizing egalitarian values in the language native to Bayesian principles. Other influential conceptualizations avoid the tradeoff conundrum (Dwork et al. 2011). Nonetheless, disparities in PPVs, FPRs, or FNRs have attracted attention in academic studies (Moss-Racusin et al., 2012; Correll et al. 2007) and the public consciousness (Angwin et al., 2016), raising the question of how to equate these constructs across different social groups. And as explained above, tradeoffs arise, which requires some way of prioritizing the three conceptions of egalitarianism.

This question has applied implications. Which conception of egalitarianism should AI developers privilege? Take, for example, an algorithm that flags posts on social media as misinformation or not. Assuming base rate differences in how frequently liberals vs. conservatives post this content (Guess, Nagler, & Tucker, 2019), there would have to be at least one disparity in PPVs, FPRs, or FNRs. As policymakers consider regulatory measures and as consumers grow more aware of these types of issues, assessing people’s relative preferences for these three conceptions of egalitarianism becomes increasingly important.

There is also theoretical value to be gained by pursuing this research. Notions of fairness have been widely studied across the social sciences (e.g., Pager, 2003), and this agenda allows for fairness to be precisely defined in mathematical terms to see which definition takes the strongest hold within the human mind. Furthermore, it is possible that people’s preferences may not be stable. In one domain, parity in FPRs may dominate, while in another domain, it is parity in PPVs that people care most about. Attitude stability and malleability has long been of interest to psychologists; here is another possible testing ground.
A methodological challenge would be eliciting preferences without bogging participants down in the math, as terms like “false positive rates” can induce an overly technical feel that might confuse more than clarify. One approach could be to implement a between-subjects design in which a different conception of egalitarianism is violated in each condition, and measures of outrage or indignation are taken. This approach is inspired by the Angwin et al. (2016) investigative report that highlighted racial disparities FPRs, which elicited strong reactions from many readers. This reaction raises the question of the following counterfactual: if the report had highlighted disparities in PPVs, would the reaction have been stronger or weaker? Experimentally capturing this relative difference could reveal what people’s preferences are for the three conceptions of egalitarianism: the more people are angered by a particular violation, the greater value they may place on it.

Base rates in the mind, base rates in the world

In many aspects of the social domain, there exist official statistics – laboriously curated over many years – that, at the very least, closely approximate the truth (e.g., gender distribution in various professions collected by the U.S. Census Bureau). In some form that varies in accuracy, these base rates enter into people’s minds: they may become exaggerated into pernicious stereotypes, or they may remain close to the state of the world. Either way, culture has placed its thumbprint upon people’s minds (Banaji, Nosek, & Greenwald, 2004).

Although this dissertation did not focus on the accuracy of people’s beliefs about base rates, these beliefs bear examination since they strongly influence subsequent mental
computations. Insofar as the base rates are inaccurate, updated beliefs will also be inaccurate, even if new information is correctly processed.

Previous research has focused on the accuracy of people’s beliefs about base rates by comparing those beliefs with census statistics (Garnham et al., 2014). As revelatory as this method can be, one shortcoming is that it is limited to areas for which there exists a reliable ground truth. Many objects, however, connote a gender even if the constructs themselves are not gendered (Palumbo, Ruta, & Bertamini, 2015). Consider the “gender” of a two drinks, whiskey vs. cosmopolitan, or of two dog breeds, pit bull vs. chihuahua. It is likely that in each pair, one is more strongly associated with men and the other with women.

A novel method to probe these base rates for which no ground truth exists is through Google Images, a search engine that returns up to thousands of relevant images for any given search term. In ongoing work, we have scraped the images from over 300 search terms and applied machine learning techniques to isolate the faces and classify the genders. From here, the proportion of women depicted for each search result can be computed. But instead of presenting proportions, these findings can be presented as face averages, which are composites of thousands of faces from each search term. For example, virtually 100% of the faces shown in the results for search “man” are men, so the corresponding face average should be highly masculine. Displayed in Figure 2 are the face averages for a few gendered search terms for which no ground truth exists per se or at least is not easily ascertainable.
Fig. 2. Face averages from Google Image search results of search terms for which no ground truth regarding gender distribution exists, or at least is not easily ascertainable.

“person drinking whiskey”

“person drinking cosmopolitan”

“pitbull owner”

“chihuahua owner”
Consistent with societal stereotypes about the gender of people who consume certain alcoholic beverages or own certain dog breeds, one face average in each pair appears more masculine and the other appears more feminine. These same techniques can be applied to search terms for which there is a known ground truth, for example “doctor” and “nurse” (see Figure 3 below).

Fig. 3. Face averages from Google Image search results for the search terms “doctor” and “nurse.”

While the face average for “doctor” is more masculine and the face composite for “nurse” is more feminine, some may see the “nurse” average as more feminine than they see the “doctor” average as masculine. This perception would be consistent with the base rate that while more women are becoming doctors, more men are not becoming nurses (Kaiser Family Foundation, 2018). Insofar as the gender distribution in image search results for various
professions track with changes, or lack thereof, in actual changes in base rates, then Google Images might be a proxy for the social environment.

The data collected from Google Images can also be used to triangulate with people’s beliefs and desires about various base rates. Descriptively, what do people believe the proportion of men vs. women to be in various domains? Prescriptively, what do people want these proportions to be? And how do each of these answers compare with the proportions computed from Google Images?

The face averages can also form open dataset that can be used to test other social perceptions concerning traits like trustworthiness or dominance. The human face is a rich source of information about a person’s states or traits, even after a brief exposure (Willis & Todorov, 2006). Testing how these perceptions relate to the original search terms could open a new window into people’s mental representations that would complement existing methods like reverse correlation (Brinkman, Todorov, & Dotsch, 2017). Given the salience of visual imagery, using Google Images as a vector into studying people’s beliefs about and desires for their social environment could be instructive.

Coda

Over the last half-century, there have been large shifts in people’s beliefs about what constitutes fair conduct. Endorsement of egalitarian values has increased, and many consider the use of group membership to be unacceptable in forming judgments about a particular individual. However, these very same people also employ base rates, thereby violating their own beliefs. Over the next half-century, another type of shift will likely take place. As
algorithms enter into domains that were once under exclusively human control, questions about how these systems operate will emerge. What data are acceptable to use? How should certain groups be portrayed? Which conception of fairness is privileged? Researching these questions in the human mind will be integral to providing guidance for a fast-changing world.


Willis, J., & Todorov, A. First impressions: Making up your mind after a 100-ms exposure ot a face. *Psychological Science, 17*(7), 592 – 598.


Wyer, N.A. (2016). Easier done than undone...by some of the people, some of the time: The role of elaboration in explicit and implicit group preferences. *Journal of Experimental Social Psychology, 63*, 77 – 85.

Detailed Materials and Methods: Experiment 1

Participants. Participants were volunteer visitors to Project Implicit (implicit.harvard.edu). Of the 599 participants who completed the procedure, 19 were excluded for going faster than 300 ms on more than 10% of trials of one IAT or both, in accordance with Greenwald, Nosek, & Banaji’s (37) scoring algorithm. An additional 6 participants were excluded for not answering both Likert-type items. The final sample consisted of 574 participants (376 females, 196 males, 2 unspecified; $M_{age} = 39.90$ years, $SD = 14.07$).

Procedure. The experiment proceeded in three parts.

1. **Measuring beliefs before the facts.** Participants read the following, which introduced them to Jonathan and Elizabeth.

   “We would like to introduce you to Jonathan and Elizabeth. Please read the following information carefully so that you can form an impression of each individual.

   Jonathan and Elizabeth are residents in a large city in the United States. Recent economic downturn has hit this city particularly hard, causing many individuals to lose their jobs. However, neither Jonathan nor Elizabeth has been affected. Both individuals are gainfully employed in their respective jobs.

   One of these individuals is employed as a doctor. The other individual is employed as a nurse.

   At this point, you do not know which person has which job. Jonathan could be the doctor and Elizabeth could be the nurse. Or, Jonathan could be the nurse and Elizabeth could be the doctor.”

   One Likert-type item measured which explicit belief they most agreed with (-3 = Jonathan is definitely the doctor, 0 = Both individuals are equally likely to be the doctor, 3 = Elizabeth is definitely the doctor). By asking for agreement, we assessed explicit beliefs in a neutral manner that did not nudge participants towards either the base rate principle or fairness principle.

   Implicit beliefs were measured with an IAT, which required participants to categorize Jonathan nicknames (Jonathan, jon, John, jonny), Elizabeth nicknames (Elizabeth, ell, Ella, eliza), Doctor words (Doctor, Medical Doctor, M.D., Physician), and Nurse words (Nurse, Registered Nurse, R.N., Nursing), as quickly as possible. Nicknames for Jonathan and Elizabeth began with upper and lowercase first letters to prevent participants from using identical visual features in the first letter to categorize each name. In one critical block, Jonathan and Doctor shared one response key and Elizabeth and Nurse shared
another response key. In the second critical block, these pairings were reversed: Elizabeth and Doctor shared one response key and Jonathan and Nurse shared the other. The order in which participants completed these blocks was counterbalanced.

2. **Learning individuating facts about Jonathan and Elizabeth.** Participants were randomly presented with one of the following sets of facts.

   (a) Control facts:

   “Jonathan, who likes to ski, vacationed in Colorado. He spent several days at a mountain resort where he skied with friends. To take a break from skiing, Jonathan and his friends relaxed in a lodge, enjoying a warm fire. Jonathan also explored a nearby city, which had a museum and several excellent restaurants. He enjoyed seeing the exhibits at the museum, and his favorite meal was at an Italian restaurant. When asked what the best part of his vacation was, Jonathan replied that it was skiing with his friends.”

   “Elizabeth, a fan of surfing, vacationed in California. She spent several days at a beach resort where she surfed with friends. To take a break from surfing, Elizabeth and her friends relaxed on the beach, enjoying the warm sun. Elizabeth also explored a nearby city, which had a zoo and several excellent restaurants. She enjoyed seeing the animals at the zoo, and her favorite meal was at French restaurant. When asked what the best part of her vacation was, Elizabeth replied that it was surfing with her friends.”

   (b) Stereotyptic facts:

   “Jonathan is a doctor at a city hospital where he specializes in emergency medicine. As a physician working in an emergency room, Jonathan cares for patients who arrive at the hospital requiring immediate medical attention. Since patients arrive with a variety of ailments, Jonathan is trained in resuscitation, cardiac life support, airway management, and some surgical procedures. After stabilizing patients, Jonathan decides whether to release them or admit them to the hospital for further treatment.”

   “Elizabeth is nurse at a retirement home where she provides care to the elderly. As a nurse working at a retirement home, Elizabeth cares for the elderly who get sick or hurt unexpectedly and for the elderly with chronic ailments. Given these different needs, Elizabeth is trained to administer medications, bandage wounds, and monitor blood pressure, heart rates and respiration. After treating the elderly, Elizabeth decides whether to release them or send them to a hospital for additional care.”
(c) Counterstereotypic facts:

“Jonathan is nurse at a retirement home where he provides care to the elderly. As a nurse working at a retirement home, Jonathan cares for the elderly who get sick or hurt unexpectedly and for the elderly with chronic ailments. Given these different needs, Jonathan is trained to administer medications, bandage wounds, and monitor blood pressure, heart rates and respiration. After treating the elderly, Jonathan decides whether to release them or send them to a hospital for additional care.”

“Elizabeth is a doctor at a city hospital where she specializes in emergency medicine. As a physician working in an emergency room, Elizabeth cares for patients who arrive at the hospital requiring immediate medical attention. Since patients arrive with a variety of ailments, Elizabeth is trained in resuscitation, cardiac life support, airway management, and some surgical procedures. After stabilizing patients, Elizabeth decides whether to release them or admit them to the hospital for further treatment.”

3. **Measuring beliefs after the facts.** The same Likert-type item and IAT from part 1 were administered again. The order in which participants completed the critical blocks in the first IAT was the same in the second IAT.

**IAT Scoring Procedure.** Following Greenwald, Nosek, & Banaji (37), we calculated two IAT D scores for each participant, one indicating an implicit belief before the facts were learned and the second indicating an implicit belief after the facts were learned. D scores were calculated such that negative values indicate a belief that Jonathan is the doctor (and Elizabeth is the nurse), whereas positive values indicate a belief that Elizabeth is the doctor (and Jonathan is the nurse). A D score of zero – which results from equal response latencies to trials in both critical blocks – indicates that both individuals are equally likely to be the doctor or nurse.

**Analyses.** Analyses for all experiments were conducted using R statistical computing’s nlme package with maximum-likelihood estimation. For both explicit and implicit beliefs, we included the interaction between time of measurement (before vs. after) and individuating facts (control vs. stereotypic vs. counterstereotypic) as a fixed effect and time of measurement nested within participant as a random effect. No other variables were included.
Detailed Materials and Methods: Experiment 2

Participants. Participants were volunteer visitors to Project Implicit (implicit.harvard.edu). Of the 848 participants who completed the procedure, 25 were excluded for going faster than 300 ms on more than 10% of trials of one IAT or both, in accordance with Greenwald, Nosek, & Banaji’s (37) scoring algorithm. An additional 15 participants were excluded for not answering both Likert-type items. The final sample consisted of 808 participants (458 females, 335 males, 15 unspecified; $M_{age} = 38.18$ years, $SD = 14.01$).

Procedure. Experiment 2 was identical to Experiment 1 except the stimuli were adapted for Richard, Jennifer, scientist, and artist. Participants were introduced to Richard and Jennifer and were told that one person is a scientist and the other is an artist. The introduction was the same as it was in Experiment 1 except the names and professions were changed. Participants expressed their explicit beliefs by indicating which statement they most agreed with ($-3 = \text{Richard is definitely the scientist}, 0 = \text{Both individuals are equally likely to be the scientist}, 3 = \text{Jennifer is definitely the scientist}$). Implicit beliefs were measured with an IAT that consisted of Richard nicknames (Richard, richy, Rick, rich), Jennifer nicknames (Jennifer, jenny, Jenn, jen), science words (Engineering, Biology, Chemistry, Physics), and arts words (Painting, Sculpture, Photography, and Drawing). The control facts were the same as in Experiment 1 except the names Jonathan and Elizabeth were replaced with Richard and Jennifer, respectively (along with the correct pronouns). Stereotypic and counterstereotypic facts were:

(a) Stereotypic facts:

“Richard is a scientist at a company called AmeriTech where he conducts research. Richard formulates hypotheses, runs experimental studies, and analyzes the data. When he discovers a noteworthy finding, Richard presents his work at scientific conferences and publishes it in scientific journals. Currently Richard is working on a promising project that could turn into an exciting scientific finding.”

“Jennifer is an artist at a company called NovaDesign where she creates artwork. Jennifer observes her surroundings, develops creative ideas, and turns those ideas into visual representations. When she creates a noteworthy work of art, Jennifer shares her work with galleries and sells it to collectors. Currently, Jennifer is working on a promising project that could turn into an exciting work of art.”
(b) Counterstereotypic facts:

“Richard is an artist at a company called NovaDesign where he creates artwork. Richard observes his surroundings, develops creative ideas, and turns those ideas into visual representations. When he creates a noteworthy work of art, Richard shares his work with galleries and sells it to collectors. Currently, Richard is working on a promising project that could turn into an exciting work of art.”

“Jennifer is a scientist at a company called AmeriTech where she conducts research. Jennifer formulates hypotheses, runs experimental studies, and analyzes the data. When she discovers a noteworthy finding, Jennifer presents her work at scientific conferences and publishes it in scientific journals. Currently Jennifer is working on a promising project that could turn into an exciting scientific finding.”
Detailed Materials and Methods: Experiment 3

**Participants.** Participants were volunteer visitors to Project Implicit (implicit.harvard.edu). Of the 719 participants who completed the procedure, 21 were excluded for going faster than 300 ms on more than 10% of trials of one IAT or both, in accordance with Greenwald, Nosek, & Banaji’s (37) scoring algorithm. An additional 7 participants were excluded for not answering both Likert-type items, and 32 participants were excluded for not correctly answering all four questions about Lapper and Affina’s genders. The final sample consisted of 659 participants (445 females, 205 males, 9 unspecified; $M_{age} = 38.33$ years, $SD = 16.73$).

**Procedure.** Experiment 3 was identical to Experiment 1 except that Lapper and Affina’s genders were taught first and then tested since their names are novel. Before being told that between Lapper and Affina, one person is a doctor and the other is a nurse, participants learned that Lapper is a man and Affina is a woman and that Lapper and Affina go by many nicknames, which were used as stimuli on the IAT. Specifically, participants were taught:

“We’d like to introduce you to two individuals, **Lapper** and **Affina**. Lapper is the individual on the left. Affina is the individual on the right. Please read the following information carefully so that you can form an impression of each individual.

As you can see in the picture, **Lapper** is a man and **Affina** is a woman.

You’ll also note that “Lapper” sounds like a man’s name and “Affina” sounds like a woman’s name.

Lapper goes by many nicknames. In addition to calling him Lapper, his friends call him Lapp, Lappster, and Pappster.

Affina also goes by many nicknames. In addition to calling her Affina, her friends call her Aff, Affie, and Fiffie.”

Next, participants answered four questions to make they understood Lapper and Affina’s genders and nicknames. The four questions are below, and the answer choices for each were Man vs. Woman.

1) Is Lapper a man or a woman?
2) Is Affina a man or a woman?
3) Affina is sometimes called Fiffie by her friends. Is Fiffie a man or a woman?
4) Lapper is sometimes called Pappster by his friends. Is Pappster a man or a woman?

The rest of the procedure was identical to that in Experiment 1. We excluded 32 participants for not answering all four of the above questions correctly. But including these 32 participants in the analyses does not change the results.

**Detailed Materials and Methods: Experiment 4**

**Participants.** Participants were volunteer visitors to Project Implicit (implicit.harvard.edu). Of the 1497 participants who completed the procedure, 59 were excluded for going faster than 300 ms on more than 10% of trials of one IAT or both, in accordance with Greenwald, Nosek, & Banaji’s (37) scoring algorithm. An additional 21 participants were excluded for not answering both Likert-type items. The final sample consisted of 1417 participants (940 females, 458 males, 19 unspecified; $M_{age} = 31.09$ years, $SD = 13.53$).

**Procedure.** Experiment 4 was identical to Experiment 2 except the stimuli were adapted for Matthew (Matthew, matty, Matt, matt) and Benjamin (Benjamin, benny, Ben, ben). All materials were the same except for the change in names and pronouns.
Supplemental Figures

**Fig. S1** Experiment 1 ($N = 574$): Distributions of explicit and implicit beliefs before facts about Jonathan and Elizabeth’s professions were learned. The dashed vertical lines indicate beliefs consistent with the fairness principle. Negative values indicate beliefs consistent with the base rate principle.

**Fig. S2** Experiment 1 ($N = 574$): Mean explicit beliefs about Jonathan and Elizabeth’s professions. Choices on the Likert-type scale are on the y-axis. Error bars are 95% CIs.
Fig. S3 Experiment 2 ($N = 808$): Distributions of explicit and implicit beliefs before facts about Richard and Jennifer’s professions were learned. The dashed vertical lines indicate beliefs consistent with the fairness principle. Negative values indicate beliefs consistent with the base rate principle.

Fig. S4 Experiment 2 ($N = 808$): Mean explicit beliefs about Richard and Jennifer’s professions. Choices on the Likert-type scale are on the y-axis. Error bars are 95% CIs.
Fig. S5 Experiment 2 ($N = 808$): Mean implicit beliefs about Richard and Jennifer’s professions. IAT $D$ scores are on the y-axis. Error bars are 95% CIs.

Fig. S6 Experiment 2: Mean implicit beliefs about Richard and Jennifer’s professions among only participants whose implicit beliefs before learning the facts were neutral ($n = 172$). IAT $D$ scores are on the y-axis. Error bars are 95% CIs.
**Fig. S7** Experiment 3 ($N = 659$): Distributions of explicit and implicit beliefs before facts about Lapper and Affina’s professions were learned. The dashed vertical lines indicate beliefs consistent with the fairness principle. Negative values indicate beliefs consistent with the base rate principle.

![Explicit beliefs](chart1)

![Implicit beliefs](chart2)

**Fig. S8** Experiment 3 ($N = 659$): Mean explicit beliefs about Lapper and Affina’s professions. Choices on the Likert-type scale are on the y-axis. Error bars are 95% CIs.
**Fig. S9** Experiment 3 ($N = 659$): Mean implicit beliefs about Lapper and Affina’s professions. IAT $D$ scores are on the y-axis. Error bars are 95% CIs.

**Fig. S10** Experiment 3: Mean implicit beliefs about Lapper and Affina’s professions among only participants whose implicit beliefs before learning the facts were neutral ($n = 143$). IAT $D$ scores are on the y-axis. Error bars are 95% CIs.
**Fig. S11** Experiment 4 ($N = 1417$): Distributions of explicit and implicit beliefs before facts about Matthew and Benjamin’s professions were learned.

**Fig. S12** Experiment 4 ($N = 1417$): Mean explicit beliefs about Matthew and Benjamin’s professions. Choices on the Likert-type scale are on the y-axis. Error bars are 95% CIs.
**Fig. S13** Experiment 4 ($N = 1417$): Mean implicit beliefs about Matthew and Benjamin’s professions. IAT $D$ scores are on the $y$-axis. Error bars are 95% CIs.
**Table S1.** Study 2. Item means and standard errors.

<table>
<thead>
<tr>
<th>Item</th>
<th>Butcher Mean (SE)</th>
<th>Firefighter Mean (SE)</th>
<th>Construction Worker Mean (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fair</td>
<td>3.28 (0.13)</td>
<td>3.19 (0.13)</td>
<td>3.62 (0.13)</td>
</tr>
<tr>
<td>Just</td>
<td>3.35 (0.12)</td>
<td>3.21 (0.12)</td>
<td>3.74 (0.13)</td>
</tr>
<tr>
<td>Accurate</td>
<td>3.70 (0.14)</td>
<td>4.09 (0.15)</td>
<td>4.09 (0.13)</td>
</tr>
<tr>
<td>Intelligent</td>
<td>3.59 (0.12)</td>
<td>3.67 (0.12)</td>
<td>3.85 (0.12)</td>
</tr>
</tbody>
</table>

**Table S2.** Study 3. Means and standard errors (in parentheses) of all 4 Likert-type, split by whether Person X offered the Bayesian judgment (Man more likely, \(n = 202\)) or egalitarian judgment (Equally likely, \(n = 200\)).

<table>
<thead>
<tr>
<th>Item</th>
<th>Man more likely Mean (SE)</th>
<th>Equally likely Mean (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fair</td>
<td>2.34 (0.11)</td>
<td>6.51 (0.07)</td>
</tr>
<tr>
<td>Just</td>
<td>2.51 (0.11)</td>
<td>6.39 (0.08)</td>
</tr>
<tr>
<td>Accurate</td>
<td>2.38 (0.12)</td>
<td>6.21 (0.11)</td>
</tr>
<tr>
<td>Intelligent</td>
<td>2.73 (0.11)</td>
<td>6.23 (0.09)</td>
</tr>
</tbody>
</table>

**Fig. S1.** Study 3. Distributions of evaluations of Person X.

1 = Unfair ... 7 = Fair
1 = Inaccurate ... 7 = Accurate
1 = Unjust ... 7 = Just
1 = Unintelligent ... 7 = Intelligent
Additional study: costly punishment
This study used another economic game to test the behavioral implications of negatively evaluating Person X. Instead of transferring money to Person X, participants had the opportunity to punish Person X, although at a financial cost to themselves.

Procedure. Four hundred thirty participants were recruited from Amazon Mechanical Turk. Each participant was compensated $0.21 and could have earned up to $0.30 more. Twenty-nine participants were excluded for not completing the procedure. The final sample consisted of 401 participants (M<sub>age</sub> = 33.87 years, SD = 10.52; 166 males, 231 females, 4 unspecified).

The procedure was identical to the procedure in Study 3 of the main text except for the financial decision participants made. Each participant was endowed with $0.30 and could give up anywhere between $0.00 and $0.10 to punish Person X, who was also endowed with $0.30 and made either the Bayesian judgment or the egalitarian judgment. For each $0.01 given up, a participant could reduce Person X’s endowment by $0.03. Thus, by giving up the maximum of $0.10, a participant could entirely take away Person X’s endowment. Participants kept the money they chose not to give up to punish Person X, and two randomly selected participants from Study 3 in the main text, one who agreed with the Bayesian judgment and another who agreed with egalitarian judgment, received the endowment amounts, less the money participants chose to deduct through costly punishment.

Results. Before discussing the amounts of money participants chose to give up to punish Person X, we first present replications of previous results. As observed previously, a majority of participants, 89%, agreed with the egalitarian judgment that the two percentages are the same. Six percent agreed with the Bayesian judgment that the two percentages differ in that the man is more likely to be the doctor, and 5% agreed that the two percentages differ in that the woman is more likely to be the doctor.

Further replicating previous results, Person X was viewed as unfair, unjust, inaccurate, and unintelligent (see Table S3 for item means and SEs) when the Bayesian judgment was offered, as indicated by means below the midpoint of 4 on the 1 to 7 Likert-type scales, Cronbach’s α = 0.93, M<sub>composite</sub> = 2.49, SE = 0.10, one-sample t(198) = -14.71, P < 0.0001, Cohen’s d = 1.04, 95% CI = [0.84, 1.31]. This effect was reversed (Fig. S2) when Person X offered the egalitarian judgment: this version of Person X was viewed as fair, just, accurate, and intelligent, Cronbach’s α = 0.85, M<sub>composite</sub> = 6.36, SD = 0.07, one-sample t(202) = 36.28, P < 0.0001, Cohen’s d = 2.55, 95% CI = [2.13, 3.13].

Critically, participants gave up more money to punishment Person X when the Bayesian judgment was offered, M = $0.02, SE = $0.002, compared to when the egalitarian judgment was offered, M = $0.004, SE = $0.001, b = $0.014, t(289.69) = -5.07, P < 0.0001, Cohen’s d = 0.51, 95% CI = [0.31, 0.71]. Also telling are the distributions of monies given up (Fig. S3). When Person X offered the egalitarian judgment, 94% chose not to give up any money to punish and only 6% engaged in costly punishment. But when Person X offered the Bayesian judgment, 28% engaged in costly punishment and 72% chose not to give up any money. These results further indicate that there are behavioral implications for negatively evaluating Person X.
Table S3. Additional study: costly punishment. Means and standard errors (in parentheses) of all 4 Likert-type items, split by whether Person X offered the Bayesian judgment (Man more likely, \( n = 198 \)) or egalitarian judgment (Equally likely, \( n = 203 \)).

<table>
<thead>
<tr>
<th></th>
<th>Man more likely</th>
<th>Equally likely</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fair</td>
<td>2.38 (0.11)</td>
<td>6.51 (0.07)</td>
</tr>
<tr>
<td>Just</td>
<td>2.39 (0.11)</td>
<td>6.47 (0.07)</td>
</tr>
<tr>
<td>Accurate</td>
<td>2.42 (0.12)</td>
<td>6.28 (0.09)</td>
</tr>
<tr>
<td>Intelligent</td>
<td>2.77 (0.11)</td>
<td>6.18 (0.08)</td>
</tr>
</tbody>
</table>

Fig. S2. Additional study: costly punishment. Distributions of evaluations of Person X.
**Fig. S3.** Additional study: costly punishment. Average amounts given up to punish Person X. Errors bars are 95% CIs. Violin plots display the distribution of amounts given up in each condition.

Money given up to punish Person X

(Max. punishment) $0.10

(No punishment) $0.00

Equally likely
n = 203

Man more likely
n = 198

Person X's answer
Additional study showing that effects are not due to the phrase “more likely” or “less likely”

This study was a stronger test of negative evaluations of Person X. In previous studies, Person X said, “the man is more likely to be a doctor.” The phrase “more likely” may imply a larger gap than the 8-percentage point difference observed in Study 4 of the main text. By revising the statement to “…8 percentage points more likely,” this study tests if negative evaluations will still emerge.

Procedure. Two hundred participants \((M_{\text{age}} = 34.00 \text{ years}, SD = 10.04; 105 \text{ males, 95 females})\) were recruited from Amazon Mechanical and compensated $0.21 each. The procedure was identical to the procedure in Study 1 of the main text, except the phrase “…more likely…” was replaced with “8 percentage points more likely”, and “…less likely…” was replaced with “8 percentage points less likely”.

Results. Once again, the majority of participants, 90%, agreed with the egalitarian judgment that the man and woman are equally likely to be a doctor. Ten percent agreed with the Bayesian judgment that the man is more likely to be a doctor; one participant agreed that the woman is more likely to be a doctor. As before, participants negatively evaluated Person X, who was viewed as unfair, \(M = 3.19, SE = 0.12\), unjust, \(M = 3.16, SE = 0.11\), inaccurate, \(M = 3.50, SE = 0.13\), and unintelligent, \(M = 3.42, SE = 0.12\), for making a quantified Bayesian judgment, as indicated by means below the midpoint of 4 on the 1-7 Likert-type scales, Cronbach’s \(\alpha = 0.93\), \(M_{\text{composite}} = 3.32, SD = 0.11\), \(t(199) = -6.20, P < 0.0001\), Cohen’s \(d = 0.44\), 95% CI = [0.29, 0.59].
Additional study showing a conceptual replication of negative evaluations of Person X

Procedure. Four hundred participants ($M_{\text{age}} = 34.68$ years, $SD = 10.89$; 150 males, 250 females) were recruited from Amazon Mechanical Turk and compensated $0.21$ each. The procedure was identical to the procedure in Study 3 in the main text except for the following differences: 1) the professions were pilot and flight attendant instead of doctor and nurse, 2) both the man and the woman communicated with air traffic control during a flight, a behavior that is highly diagnostic of being the pilot, and 3) there was no economic game.

Participants were instructed to imagine a man and a woman who both work for the same airline. One person is a pilot and the other person is a flight attendant. But who is the pilot vs. flight attendant is unknown. In counterbalanced order, participants were instructed to assume that the man had communicated with air traffic control during a flight, in which case the probability that the man is the pilot is an unknown percentage. Participants were then instructed to assume that the woman had communicated with air traffic control during a flight, in which case the probability that the woman is the pilot is another unknown percentage. Participants indicated whether they agreed that a) the two percentages differ in that the man is less likely to be the pilot, b) the two percentages are equivalent, or c) the two percentages differ in that the man is more likely to be the pilot. As before, the order in which the man and woman were compared was randomly assigned.

Participants then read about Person X, who, after learning the same information as participants, offered either the Bayesian judgment or egalitarian judgment, based on random assignment. Participants then evaluated how fair, just, accurate, and intelligent Person X’s statement was on four Likert-type scales that each ranged from 1 to 7 (e.g., 1 = Extremely unfair ... 7 = Extremely fair) before providing open-ended text responses of their impressions of Person X.

Results. The results were replicated. A majority of participants, 81%, agreed with the egalitarian judgment that the two percentages are equivalent, whereas 15% agreed with the Bayesian judgment that the two percentages differ in that the man is more likely to be the pilot, and 4% agreed that the two percentages differ in that the man is less likely to be the pilot.

Further replicating previous results, Person X was viewed as unfair, unjust, inaccurate, and unintelligent (see Table S4 for items means and SEs) when the Bayesian judgment was offered, as indicated by means below the midpoint of 4 on the 1 to 7 Likert-type scales, Cronbach’s $\alpha = 0.93$, $M_{\text{composite}} = 3.11$, $SD = 0.11$, one-sample $t(198) = -7.99$, $P < 0.0001$, Cohen’s $d = 0.57$, 95% CI = [0.41, 0.74]. This effect was reversed (Fig. S4) 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, Cronbach’s $\alpha = 0.85$, $M_{\text{composite}} = 6.30$, $SD = 0.06$, one-sample $t(200) = 36.60$, $P < 0.0001$, Cohen’s $d = 2.59$, 95% CI = [2.11, 3.27].
Table S4. Additional study showing conceptual replication negative evaluations of Person X. Means and standard errors (in parentheses) of all 4 Likert-type items, split by whether Person X offered the Bayesian judgment (Man more likely, $n = 199$) or egalitarian judgment (Equally likely, $n = 201$).

<table>
<thead>
<tr>
<th></th>
<th>Man more likely</th>
<th>Equally likely</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fair</td>
<td>2.96 (0.12)</td>
<td>6.55 (0.06)</td>
</tr>
<tr>
<td>Just</td>
<td>2.96 (0.12)</td>
<td>6.43 (0.06)</td>
</tr>
<tr>
<td>Accurate</td>
<td>3.15 (0.13)</td>
<td>6.05 (0.07)</td>
</tr>
<tr>
<td>Intelligent</td>
<td>3.37 (0.12)</td>
<td>6.18 (0.07)</td>
</tr>
</tbody>
</table>

Fig. S4. Additional study showing conceptual replication negative evaluations of Person X. Distributions of evaluations of Person X.
**Fig. S5.** Study 4: sponge bath conditions. **A.** Minimal differences in likelihood ratios were observed between participants who learned that the man vs. woman had given a sponge bath to a patient, Median\textsubscript{Man} = -2.36 vs. Median\textsubscript{Woman} = -2.77, Wilcoxon $P = 0.05$, $r = 0.11$. Moreover, the log of these likelihood ratios were less than zero, indicating that giving a sponge bath is diagnostic of who is the nurse (i.e., not the doctor). **B.** Because priors favored the man to be the doctor and because the data were diagnostic of the profession nurse, the probability that each target was the doctor was low. However, model posteriors still favored the man to be the doctor, $M_{\text{Model Posterior, Man}} = 24.3\%$ vs. $M_{\text{Model Posterior, Woman}} = 7.4\%$, $b = 0.17$, $t(890) = 7.54$, $P < 0.0001$, $r = 0.25$. This disparity was also observed among these participants’ reported posteriors, $M_{\text{Reported Posterior, Man}} = 31.6\%$ vs. $M_{\text{Reported Posterior, Woman}} = 13.1\%$, $b = 0.19$, $t(890) = 8.28$, $P < 0.0001$, $r = 0.27$. Furthermore, relatively small differences were observed between model and reported posteriors among participants who learned that the man had given a sponge bath, $M_{\text{Model Posterior, Man}} = 24.3\%$ vs. $M_{\text{Reported Posterior, Man}} = 31.6\%$, $b = -0.07$, $t(1780) = -4.08$, $P < 0.0001$, $r = 0.10$, and among participants who learned that the woman had given a sponge bath, $M_{\text{Model Posterior, Woman}} = 7.4\%$ vs. $M_{\text{Reported Posterior, Woman}} = 13.1\%$, $b = -0.06$, $t(1780) = -3.02$, $P = 0.003$, $r = 0.07$. Error bars are 95% CIs.
Fig. S6. Study 4: CPR conditions. A. Minimal differences in likelihood ratios were observed between participants who learned that the man vs. woman had performed CPR, $\text{Median}_{\text{Man}} = -0.19$ vs. $\text{Median}_{\text{Woman}} = -0.18$, Wilcoxon $P = 0.66$, $r = 0.03$. Moreover, the log of these likelihood ratios were close to zero, indicating that performing CPR is relatively non-diagnostic of who is the doctor. B. Because priors favored the man to be the doctor and because the data were relatively non-diagnostic, model posteriors remained close to priors. Reported posteriors were relatively similar to model posteriors, $t_{(1780)} < |3.44|$, $Ps > 0.0006$, $rs < 0.09$. Error bars are 95% CIs.
Fig. S7. Study 4: surgery conditions. A. The correspondence between model and reported posteriors is present at the level of the individual participant. By subtracting each participant’s model posterior from his or her reported posterior, we calculate an accuracy score for each participant, with zero being completely accurate. The distribution of these accuracy scores is shown below. The mode of this distribution is zero, which suggests the statistical savvy of the individual rather than a wisdom of the crowds effect. B. Unlike the representativeness heuristic, the Bayesian account predicts that participants’ reported posteriors are directly proportional to their likelihood estimates. This positive relationship emerges among participants with non-infinite likelihoods, $r = 0.30, P < 0.0001$, and remains when controlling for participants’ priors, $B = 0.25, t(189) = 3.67, P = 0.0003, r = 0.26$. 

![Graph A: Proportion of participants vs. Reported Posterior - Model Posterior](image1)

![Graph B: Log Likelihood Ratio vs. Reported Posterior](image2)
Fig. S8. Study 4: surgery conditions. The statistical significance of the four critical comparisons is robust to the choice of adjustment factor when participants’ probability judgments are logit transformed. The adjustment factor is necessary to avoid logit transforming probabilities of 0 or 1. Each panel shows one of the critical comparisons in the surgery conditions, and the $P$ value is plotted as a function of the adjustment factor. For all comparisons except for one, whether $P$ is greater or less than 0.05 (red horizontal line) does not depend on the adjustment factor.
Additional study that conceptually replicates Bayesian judgments
This study assessed probability judgments in the domain of pilot vs. flight attendant.

Procedure. Nine hundred sixty four participants were recruited from Amazon Mechanical Turk and compensated $0.50 each. Nineteen participants were excluded because they provided priors that cannot be updated according to Bayes’ rule, and six participants were excluded because their model posteriors could not be computed since they answered 0% to both likelihood questions. Another six participants indicated they had looked up answers to some of the questions in the study, but these participants are retained in the analyses (conclusions do not change based on whether these participants are included or excluded). While 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 in the main text. The final sample consisted of 939 participants (M_{age} = 36.59 years, SD = 12.25; 426 males, 510 females, 3 unspecified).

The procedure consisted of the same three parts as Study 4. Participants provided their subjective priors about who was the pilot vs. flight attendant and were randomly assigned to learn one of following six pieces of data, after which they provided their subjective posteriors.

i. The man communicated with air traffic control during a flight.
ii. The woman communicated with air traffic control during a flight.
iii. The man beverages to passengers during a flight.
iv. The woman served beverages to passengers during a flight.
vi. The man went through a special line at airport security.
vi. The woman went through a special line at airport security.

Communicating with air traffic control was chosen because it is highly diagnostic of the person being a pilot. Serving beverages was chosen because it is highly diagnostic of the person being a flight attendant. Going through a special line at airport security was chosen because it is relatively non-diagnostic of profession, as both pilots and flight attendants do this. For the primary analysis, only the first two conditions (i and ii, communicated with air traffic control) are discussed. Data from the other four conditions (iii – vi) are presented in Figs. S9-S10.

Participants also estimated two likelihoods, the likelihood of observing the datum given the hypothesis that the target they learned about is the pilot as well as the likelihood of observing the datum given the hypothesis that the target they learned about is the flight attendant. If a participant learned that the woman had communicated with air traffic control during a flight, that participant estimated the percentage of female pilots who communicate with traffic control during a flight, as well as the percentage of female flight attendants who communicate with traffic control during a flight. If a participant learned that the man had communicated with air traffic control during a flight, that participant answered the same two questions except about male pilots and male flight attendants. As before in Study 4 of the main text, each participant’s priors and likelihoods were entered into Bayes’ rule to compute a model posterior, which was then compared against the posterior the participant had reported.
**Results.** When the target was a man, he was more likely to be the pilot *a priori* than when the target was a woman, $M_{\text{Man}} = 71.2\%$ vs. $M_{\text{Woman}} = 26.1\%$, $b = 0.45$, $t(933) = 18.88$, $P < 0.0001$, $r = 0.53$, as 77% of participants reported priors that favored the man over the woman to be the pilot.

Consistent with previously observed likelihood estimates, likelihood estimates in the current study reflected the fact that not everyone who communicates with air traffic control during a flight is necessarily a pilot. Regardless of the gender of the target who exhibited this behavior, the majority of participants indicated that a non-zero percentage of flight attendants also communicate with air traffic control, resulting in likelihoods less than infinity. Moreover, only a small difference in likelihoods was observed between the two conditions, $\text{Median}_{\text{Man}} = 2.23$ vs. $\text{Median}_{\text{Woman}} = 1.96$, Wilcoxon $P = 0.39$, $r = 0.05$, which suggests that participants may have found the datum of communicating with air traffic control to be equally diagnostic of being a pilot, irrespective of the target’s gender (Fig. S11A). Many participants (<24% in both conditions) found the datum 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 pilot and because likelihoods were similar between the two conditions, 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, $M_{\text{Model Posterior, Man}} = 90.8\%$ vs. $M_{\text{Model Posterior, Woman}} = 63.0\%$, $b = 0.28$, $t(933) = 11.60$, $P < 0.0001$, $r = 0.35$. As was the case in Study 2, this disparity was also observed among participants’ reported posteriors, $M_{\text{Reported Posterior, Man}} = 85.8\%$ vs. $M_{\text{Reported Posterior, Woman}} = 67.3\%$, $b = 0.18$, $t(933) = 7.73$, $P < 0.0001$, $r = 0.25$. Further replicating previous results, small differences were observed between model posteriors and reported posteriors among participants who learned that the man had communicated with air traffic control, $M_{\text{Model Posterior, Man}} = 90.8\%$ vs. $M_{\text{Reported Posterior, Man}} = 85.8\%$, $b = 0.05$, $t(1866) = 2.59$, $P = 0.01$, $r = 0.06$, and among participants who had learned that the woman had communicated with air traffic control, $M_{\text{Model Posterior, Woman}} = 63.0\%$; vs. $M_{\text{Reported Posterior, Woman}} = 67.3\%$, $b = -0.04$, $t(1866) = -2.24$, $P = 0.03$, $r = 0.05$. So once again, the posteriors reported by participants were close to the posteriors they should have reported according to Bayesian rationality (Fig. S11B).

Additional analyses 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 are robust when participants’ probability judgments are logit transformed with a wide range of adjustment factors (Figs. S12-S13). In sum, a man who communicated with air traffic control during a flight was judged more likely to be a pilot than a woman who exhibited the same behavior.
**Fig. S9.** Additional study that conceptually replicates Bayesian judgments: served beverage conditions. A. Minimal differences in likelihood ratios were observed between participants who learned that the man vs. woman had served beverages to passengers, $\text{Median}_{\text{Man}} = -\text{Inf}$ vs. $\text{Median}_{\text{Woman}} = -\text{Inf}$, Wilcoxon $P = 0.67$, $r = 0.02$. Moreover, the log of these likelihood ratios were less than zero, indicating that serving beverages is diagnostic of who is the flight attendant (i.e., not the pilot). B. Because priors favored the man to be the pilot and because the data were diagnostic of the profession flight attendant, the probability that each target was the pilot was low. However, model posteriors still favored the man to be the pilot, $M_{\text{Model Posterior, Man}} = 10.0\%$ vs. $M_{\text{Model Posterior, Woman}} = 2.6\%$; $b = 0.07$, $t(933) = 3.06$, $P = 0.002$, $r = 0.10$. This disparity was also observed among these participants’ reported posteriors, $M_{\text{Reported Posterior, Man}} = 25.9\%$ vs. $M_{\text{Reported Posterior, Woman}} = 10.8\%$; $b = 0.15$, $t(933) = 6.26$, $P < 0.0001$, $r = 0.20$. Reported posteriors were greater than model posteriors among participants who learned that the man had served beverages, $M_{\text{Model Posterior, Man}} = 10.0\%$ vs. $M_{\text{Reported Posterior, Man}} = 25.9\%$; $b = -0.16$, $t(1866) = -8.34$, $P < 0.0001$, $r = 0.19$, and among participants who learned that the woman had served beverages, $M_{\text{Model Posterior, Woman}} = 2.6\%$ vs. $M_{\text{Reported Posterior, Woman}} = 10.8\%$; $b = -0.08$, $t(1866) = -4.22$, $P < 0.0001$, $r = 0.10$. Error bars are 95% CIs.
Fig. S10. Additional study that conceptually replicates Bayesian judgments: special line conditions. **A.** Minimal differences in likelihood ratios were observed between participants who learned that the man vs. woman had gone through a special line at airport security, $\text{Median}_{\text{Man}} = 0$ vs. $\text{Median}_{\text{Woman}} = 0$, Wilcoxon $P = 0.01$, $r = 0.14$. Moreover, the log of these likelihood ratios were close to zero, indicating that going through a special line is relatively non-diagnostic of who is the pilot. **B.** Because priors favored the man to be the pilot and because the data were relatively non-diagnostic, model posteriors remained close to priors. Reported posteriors were similar to model posteriors, $ts(1866) < |3.36|$, $Ps > 0.0008$, $rs < 0.08$. Error bars are 95% CIs.
**Fig. S11.** Additional study that conceptually replicates Bayesian judgments: communicated with air traffic control (ATC) conditions. **A.** Distribution of likelihood ratios (log scaled) in each condition. **B.** Average judgments among participants in each condition. Priors indicate judgments before participants learned that the target had communicated with air traffic control. Model posteriors indicate judgments participants should make from a Bayesian perspective. Reported posteriors indicate judgments participants actually made. Error bars are 95% CIs.
Fig. S12. Additional study that conceptually replicates Bayesian judgments: communicated with air traffic control (ATC) conditions. A. The correspondence between model and reported posteriors is present at the level of the individual participant. By subtracting each participant’s model posterior from his or her reported posterior, we calculate an accuracy score for each participant, with zero being completely accurate. The distribution of these accuracy scores is shown below. The mode of this distribution is zero, which suggests the statistical savvy of the individual rather than a wisdom of the crowds effect. B. Unlike the representativeness heuristic, the Bayesian account predicts that participants’ reported posteriors are directly proportional to their likelihood estimates. This positive relationship emerges among participants with non-infinite likelihoods, \( r = 0.30, P < 0.0001 \), and remains when controlling for participants’ priors, \( B = 0.31, t(250) = 5.77, P < 0.0001, r = 0.34 \).
Fig. S13. Additional study that conceptually replicates Bayesian judgments: communicated with air traffic control (ATC) conditions. The statistical significance of the four critical comparisons is robust to the choice of adjustment factor when participants’ probability judgments are logit transformed. The adjustment factor is necessary to avoid logit transforming probabilities of 0 or 1. Each panel shows one of the critical comparisons in the communicated w/ATC conditions, and the $P$ value is plotted as a function of the adjustment factor. Whether $P$ is greater or less than 0.05 (red horizontal line) does not depend on the adjustment factor.
**Fig. S14.** Study 5. **A.** Distribution of likelihood ratios (log scaled) in each condition. **B.** Average judgments among participants in each condition. Priors indicate judgments before participants learned that the target had communicated with air traffic control. Model posteriors indicate judgments participants should make from a Bayesian perspective. Reported posteriors indicate judgments participants actually made. Error bars are 95% CIs.
Table S5: Study 5. Proportion of participants who agreed that the woman is more likely to be a doctor, conditional on both the man and woman having performed surgery, that they’re equally likely to be a doctor, or that the man is more likely to be a doctor (rows). Proportion of participants whose priors favored the woman to be the pilot, both the man and woman equally likely to be the pilot, or the man to be the pilot (columns). Joint proportions are inside the cells and marginal proportions are in the margins. Along the main diagonal are the minority of participants who were consistent by using the base rate in both parts of the study. The cell containing highest proportion of participants (70.69%) are those who used gendered base rates when making their probability judgments but not when indicating the statement they agreed with.

<table>
<thead>
<tr>
<th></th>
<th>Woman more likely to be pilot</th>
<th>Equally likely to be pilot</th>
<th>Man more likely to be pilot</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Woman more likely to be doctor</strong></td>
<td>0%</td>
<td>0.29%</td>
<td>1.15%</td>
</tr>
<tr>
<td><strong>Equally likely to be doctor</strong></td>
<td>1.15%</td>
<td>7.18%</td>
<td>70.69%</td>
</tr>
<tr>
<td><strong>Man more likely to be doctor</strong></td>
<td>0.57%</td>
<td>0.29%</td>
<td>18.68%</td>
</tr>
<tr>
<td><strong>1.72%</strong></td>
<td>7.76%</td>
<td>90.52%</td>
<td>100%</td>
</tr>
</tbody>
</table>
**Fig. S15.** Study 5. Scatterplot and line of best of fit showing the relationship between statistical accuracy on y-axis (model posterior subtracted from reported posterior) and evaluations of Person X on the x-axis (average of four Likert-type items). Distributions of each variable are in the margins. The relationship is weak, $r = -0.10$, $P = 0.06$, indicating that participants made accurate Bayesian judgments irrespective of how they evaluated Person X.
Conceptual replication of Study 5

This study was served as a conceptual replication of Study 5 by reversing the scenarios: participants made probability judgments in the doctor scenario and then evaluated Person X in the pilot scenario.

Procedure. Three hundred fifty nine participants were recruited from Amazon Mechanical Turk and compensated $0.71 each. Four participants were excluded because they provided priors that cannot be updated according to Bayes’ rule. The final sample consisted of 355 participants ($M_{age} = 35.24$ years, $SD = 11.73$; 196 males, 158 females, 1 unspecified).

The study consisted of three parts. In the first part, participants were randomly assigned to learn that either a man or woman had performed surgery. Participants provided their priors, posteriors, and likelihoods for this scenario, just as they did in Study 4 in the main text. As before, a model posterior was computed for each participant and compared to this 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 almost the identical procedure 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 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. Thus, this study reversed the doctor and pilot scenarios.

Results. Bayesian judgments were again observed, which replicates previous results (Fig. S16). Model posteriors favored the man over the woman to be the doctor even though both targets had performed surgery, $M_{Model Posterior, Man} = 85.3\%$ vs. $M_{Model Posterior, Woman} = 65.6\%$, $b = 0.20$, $t(353) = 8.27$, $P < 0.0001$, $r = 0.40$. As before, this disparity was also observed among participants’ reported posteriors, $M_{Reported Posterior, Man} = 79.7\%$ vs. $M_{Reported Posterior, Woman} = 72.4\%$, $b = 0.07$, $t(353) = 3.09$, $P = 0.002$, $r = 0.16$.

Further replicating previous results, relatively small differences were observed between model posteriors and reported posteriors among participants who learned that the man had performed surgery, $M_{Model Posterior, Man} = 85.3\%$ vs. $M_{Reported Posterior, Man} = 79.7\%$, $b = 0.06$, $t(706) = 2.90$, $P = 0.004$, $r = 0.11$, and among participants who had learned that the woman had performed surgery, $M_{Model Posterior, Woman} = 65.6\%$ vs. $M_{Reported Posterior, Woman} = 72.4\%$, $b = -0.07$, $t(706) = -3.39$, $P = 0.007$, $r = 0.13$. So once again, posteriors reported by participants were close to the posteriors they should have reported according to Bayesian rationality.

These participants who made Bayesian judgments were divided in which judgment they agreed with: 44.2% agreed with the egalitarian judgment that the man and woman are equally likely to be a pilot, conditional on both having communicated with air traffic control during a flight, 52.1% agreed with the Bayesian judgment that the man is more likely to be a pilot, and 3.7% agreed that the woman is more likely to be a pilot.

Participants, on average, made slightly positive evaluations of Person X, who was rated above the midpoint of 4 on the 1-7 Likert-type scales. Person X was viewed as fair, $M = 4.26$, $SE = 0.09$, just, $M = 4.30$, $SE = 0.09$, accurate, $M = 4.84$, $SE = 0.08$, and intelligent, $M = 4.58$, $SE = 0.08$, for making the Bayesian judgment that the man is more likely to be the pilot, Cronbach’s
\[ \alpha = 0.89, M_{\text{composite}} = 4.49, SE = 0.07, \text{one-sample } t(354) = 6.80, P < 0.0001, \text{Cohen's } d = 0.36, 95\% CI = [0.25, 0.48]. \]

These diminished effects likely stem from three sources. First, as was the case the Study 5 in the main text, the preceding statistical judgments – both the main judgments concerning the gender of the doctor and the filler judgments – made base rates more salient. Second, base rates concerning the gender distribution among pilots are stronger than the base rates concerning the gender distribution among doctors. And third, communicating with air traffic control may not be seen as diagnostic of the profession pilot as performing surgery is of the profession doctor (see proportion of infinite likelihood ratios in Study 4 of main text vs. proportion of infinite likelihood ratios in study in Supplemental Materials that conceptually replicates Study 4). Together, these three features make this study an especially conservative way of testing if the same participants make Bayesian judgments and negatively evaluate others for doing likewise.

Despite how conservative this study was, the critical analysis of regressing reported probabilities on evaluations of Person X replicates the results of Study 5 (Fig. S17). Participants judged that the man is more likely to be the doctor than the woman, regardless of their evaluation of Person X, \( F(1, 351) = 7.52, P = 0.006, \eta^2 = 0.02, 95\% CI = [0.002, 0.06] \). Even participants who were critical of Person X judged that the man is more likely to be the doctor than the woman, conditional on each having performed surgery. Statistically significant differences between reported probabilities for the man vs. woman conditions hold for participants whose average evaluation of Person X is 3.5 or higher on the 1 to 7 scale, which is the 20th percentile. Although qualitatively, even participants who were the most critical of Person X, as indicated by ratings of 1 on all four items, judged that the man is more likely to be the doctor than the woman. So even though the effects here are weaker here, they are still present.

Further consistent with Study 5, participants were equally and highly accurate irrespective of how they felt towards Person X, as evidenced by the minimal difference between their model and reported posteriors across the entire range of evaluations (Fig. S18). Thus, participants accurately judged that the man is more likely to be the doctor than a woman. These participants then proceeded to criticize Person X for making a conceptually similar Bayesian judgment.
Fig. S16. Conceptual replication of Study 5. A. Distribution of likelihood ratios (log scaled) in each condition. B. Average judgments among participants in each condition. Priors indicate judgments before participants learned that the target had communicated with air traffic control. Model posteriors indicate judgments participants should make from a Bayesian perspective. Reported posteriors indicate judgments participants actually made. Error bars are 95% CIs.
Fig. S17. Conceptual replication of Study 5. Reported posterior probabilities as a function of evaluations of Person X (average of four Likert-type items). Grey bands are SEs.
Fig. S18. Conceptual replication of Study 5. Scatterplot and line of best of fit showing the relationship between statistical accuracy on y-axis (model posterior subtracted from reported posterior) and evaluations of Person X on the x-axis (average of four Likert-type items). Distributions of each variable are in the margins. The relationship is weak, \( r = -0.03, P = 0.56 \), indicating that participants made accurate Bayesian judgments irrespective of how they evaluated Person X.
Probability calculus

Assume a Bayesian network of the form $A \rightarrow B \rightarrow C$

The joint probability distribution is:

$$P(A, B, C) = P(A) \times P(B|A) \times P(C|B)$$

Using Bayes’ rule and the joint probability distribution above, it can be shown that judgments about $C$ do not depend on $A$ once $B$ is known:

$$P(C|A, B) = \frac{P(A, B, C)}{P(A, B)} = \frac{P(A) \times P(B|A) \times P(C|B)}{P(A) \times P(B|A)} = P(C|B)$$

*When the base rate is social:*

- Let $A$ be gender (male vs. female)
- Let $B$ be profession (doctor vs. nurse)
- Let $C$ be invitation decision (invite vs. don’t invite)

$$P(C = \text{invite} \mid A = \text{male}, B = \text{doctor}) = P(C = \text{invite} \mid B = \text{doctor})$$

$$P(C = \text{invite} \mid A = \text{female}, B = \text{doctor}) = P(C = \text{invite} \mid B = \text{doctor})$$

By the transitive property:

$$P(C = \text{invite} \mid A = \text{male}, B = \text{doctor}) = P(C = \text{invite} \mid A = \text{female}, B = \text{doctor})$$

Likewise:

$$P(C = \text{invite} \mid A = \text{male}, B = \text{nurse}) = P(C = \text{invite} \mid B = \text{nurse})$$

$$P(C = \text{invite} \mid A = \text{female}, B = \text{nurse}) = P(C = \text{invite} \mid B = \text{nurse})$$

Again, by the transitive property:

$$P(C = \text{invite} \mid A = \text{male}, B = \text{nurse}) = P(C = \text{invite} \mid A = \text{female}, B = \text{nurse})$$
**Probability calculus (continued)**

*When the base rate is nonsocial:*

- Let $A$ be material (metal vs. plastic)
- Let $B$ be status (intact vs. broken)
- Let $C$ be keep decision (keep vs. don’t keep)

\[
P(C = \text{keep} \mid A = \text{metal}, B = \text{intact}) = P(C = \text{keep} \mid B = \text{intact})
\]
\[
P(C = \text{keep} \mid A = \text{plastic}, B = \text{intact}) = P(C = \text{keep} \mid B = \text{intact})
\]

By the transitive property:
\[
P(C = \text{keep} \mid A = \text{metal}, B = \text{intact}) = P(C = \text{keep} \mid A = \text{plastic}, B = \text{intact})
\]

Likewise:
\[
P(C = \text{keep} \mid A = \text{metal}, B = \text{broken}) = P(C = \text{keep} \mid B = \text{broken})
\]
\[
P(C = \text{keep} \mid A = \text{plastic}, B = \text{broken}) = P(C = \text{keep} \mid B = \text{broken})
\]

Again, by the transitive property:
\[
P(C = \text{keep} \mid A = \text{metal}, B = \text{broken}) = P(C = \text{keep} \mid A = \text{plastic}, B = \text{broken})
\]
Supplementary Study

Replication of Experiment 1 with a comprehension check that participants had to answer from memory. In this replication, the procedures and materials were identical to those in the original Experiment 1 except for the following difference: after making two judgments, each participant answered one of the following two comprehension check questions from memory, depending on whether they had been assigned to the nonsocial or social scenario. The scenario was not available for participants to reference when they answered the comprehension check, so they had to answer it from memory. The correct answers are bolded, and the order in which the answer choices appeared was randomized. A total of 602 participants were recruited from Amazon Mechanical Turk.

<table>
<thead>
<tr>
<th>Comprehension check for nonsocial scenario</th>
<th>Comprehension check for social scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Think back to the scenario you just read.</td>
<td>Think back to the scenario you just read.</td>
</tr>
<tr>
<td>Will the factory keep more intact spoons or broken spoons?</td>
<td>Will the charity ask more doctors or nurses?</td>
</tr>
<tr>
<td>• More intact spoons</td>
<td>• More doctors</td>
</tr>
<tr>
<td>• An equal number of each</td>
<td>• An equal number of each</td>
</tr>
<tr>
<td>• More broken spoons</td>
<td>• More nurses</td>
</tr>
</tbody>
</table>

High comprehension was observed for both the nonsocial and social scenarios: 77% of participants randomly assigned to the social scenario understood that the charity will invite more doctors than nurses, and 86% of participants randomly assigned to the nonsocial scenario understood that the factory will keep more intact spoons than broken spoons (Supplementary Fig. 1).

Furthermore, the same results from Experiment 1 are again observed when the minority of participants who failed the comprehension check are excluded (Supplementary Fig. 2). In the nonsocial spoon scenario, the plastic spoon was judged less likely to be kept regardless of whether the two spoons in question were broken \([M_{\text{broken metal}} = 4.85 \text{ vs. } M_{\text{broken plastic}} = 3.73; b = 1.12, t(484) = 5.97, P < 0.0001}] or intact \([M_{\text{intact metal}} = 6.76 \text{ vs. } M_{\text{intact plastic}} = 6.08; b = 0.68, t(484) = 3.64, P = 0.0003}]. But in the social gender scenario, judgments again erroneously relied upon group differences. A male nurse was judged less likely to be invited than a female nurse \([M_{\text{male nurse}} = 3.71 \text{ vs. } M_{\text{female nurse}} = 5.42; b = -1.72, t(484) = -8.91, P < 0.0001}]], but a female doctor was judged less likely to be invited than a male doctor \([M_{\text{male doctor}} = 6.29 \text{ vs. } M_{\text{female doctor}} = 4.77; b = 1.52, t(484) = 7.67, P < 0.0001}]. This replication of Experiment 1 with a comprehension check provides assurance that the effects are not due the lack of comprehension, as the same participants who comprehended the scenario also produced errors in their judgments.
**Supplementary Fig. 1.** Replication of Experiment 1 with a comprehension check that participants had to answer from memory (N = 602). Here, we plot the results of the comprehension check. In both the social and nonsocial scenarios, a clear majority of participants correctly comprehended that either more intact spoons would be kept by the factory or that more doctors would be invited by the charity.
Supplementary Fig. 2. Replication of Experiment 1 with a comprehension check that participants had to answer from memory (N = 602). Of the 602 participants recruited, 488 passed the comprehension check (81%). Only the results of these 488 participants are plotted here. Participants’ average judgments when nonsocial vs. social base rates should not have been used. Each participant was randomly assigned to make two judgments on a 1 to 7 scale (e.g., broken metal spoon vs. broken plastic spoon). All possible pairs of judgments a participant could have been randomly assigned to make are on the x-axes. Statistical tests compare the means of each pair of judgments. Points show the distribution of judgments. Error bars are 95% CIs. *** P < 0.001, ** P < 0.01, * P < 0.05, n.s. P > 0.05.
Supplementary Study

Replication of Experiment 1 with a comprehension check that participants could answer by referencing the scenario. Another replication was Experiment 1 with a comprehension check was also conducted ($N = 454$ from Amazon Mechanical Turk). However, participants in this experiment were allowed to reference the scenario while answering the comprehension check, which makes the results of this comprehension check less informative. Nonetheless, for complete disclosure, we present the results of this replication here, which once again demonstrate the erroneous judgments observed in other experiments (Supplementary Figs. 3-4).

Supplementary Fig. 3. Replication of Experiment 1 with a comprehension check that participants could answer by referencing the scenario ($N = 454$). Here, we plot the results of the comprehension check. In both the social and nonsocial scenarios, a clear majority of participants correctly comprehended that either more intact spoons would be kept by the factory or that more doctors would be invited by the charity.
Supplementary Fig. 4. Replication of Experiment 1 with a comprehension check that participants could answer by referencing the scenario ($N = 454$). Of the 454 participants recruited, 399 passed the comprehension check (87.9%). Only the results of these 399 participants are plotted here. Participants’ average judgments when nonsocial vs. social base rates should not have been used. Each participant was randomly assigned to make two judgments on a 1 to 7 scale (e.g., broken metal spoon vs. broken plastic spoon). All possible pairs of judgments a participant could have been randomly assigned to make are on the x-axes. Statistical tests compare the means of each pair of judgments. Points show the distribution of judgments. Error bars are 95% CIs. *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, n.s. $P > 0.05$. 

---

Nonsocial

Social

![Nonsocial and Social Diagrams](image-url)
Supplementary Figure

Supplementary Fig. 3. Experiment 2. Participants’ average judgments when various nonsocial base rates should not have been used. Each participant was randomly assigned to make two judgments on a 1 to 7 scale (e.g., broken metal spoon vs. broken plastic spoon). All possible pairs of judgments a participant could have been randomly assigned to make are on the x-axes. Statistical tests compare the means of each pair of judgments. Points show the distribution of judgments. Error bars are 95% CIs. *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, n.s. $P > 0.05$.
Supplementary Figure

Supplementary Fig. 4. Participants’ average judgments when various social base rates should not have been used. Each participant was randomly assigned to make two judgments on a 1 to 7 scale (e.g., male nurse vs. female nurse). All possible pairs of judgments a participant could have been randomly assigned to make are on the x-axes. Statistical tests compare the means of each pair of judgments. Points show the distribution of judgments. Error bars are 95% CIs. *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, n.s. $P > 0.05$. 
**Supplementary Fig. 5.** Experiment 2 ($N = 1,672$) collapsed results. Participants’ average judgments when various nonsocial vs. social base rates should not have been used. Pairs of judgments akin to the judgments in Experiment 1 are on the x-axes. Statistical tests compare the means of each pair of judgments. Points show the distribution of judgments. Error bars are 95% CIs. *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, n.s. $P > 0.05$. 

Various nonsocial base rates (collapsed)

Various social base rates (collapsed)
**Supplementary Figure**

**Supplementary Fig. 6.** Experiment 3 (N = 398). Participants’ average judgments when nonsocial vs. social base rates should not have been used. In this experiment, the within- and between-subject conditions were switched, and the same results were replicated. Statistical tests compare the means of each pair of judgments. Points show the distribution of judgments. Error bars are 95% CIs. *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, n.s. $P > 0.05$. 
Supplementary Study

Experiment 5. Experiment 5 was identical to the spoon and gender scenarios in Experiment 2, except the sentences that explicitly conveyed disparate outcomes were removed. Thus, Experiment 5 remedied only the first problem discussed in Experiment 4.

When only this remedy was implemented, the previous results were replicated (Supplementary Fig. 7). Violations of the Markov condition strongly depended on whether the base rates were nonsocial or social \( F(1, 398) = 61.26, P < 0.0001 \). In the spoon scenario, which contained nonsocial base rates, the plastic spoon was judged less likely to be kept regardless of whether the two spoons in question were broken \( M_{\text{broken metal}} = 4.69 \) vs. \( M_{\text{broken plastic}} = 3.57; b = 1.12, t(398) = 5.33, P < 0.0001 \) or intact \( M_{\text{intact metal}} = 6.44 \) vs. \( M_{\text{intact plastic}} = 5.49; b = 0.95, t(398) = 4.51, P < 0.0001 \).

But in the gender scenario, which contained social base rates, judgments erroneously relied upon group differences. A male nurse was judged less likely to be invited than a female nurse \( M_{\text{male nurse}} = 3.95 \) vs. \( M_{\text{female nurse}} = 5.71; b = -1.76, t(398) = -8.15, P < 0.0001 \), but a female doctor was judged less likely to be invited than a male doctor \( M_{\text{male doctor}} = 6.16 \) vs. \( M_{\text{female doctor}} = 4.78; b = 1.38, t(398) = 6.61, P < 0.0001 \). Once again, social judgments broke with Bayesian rationality and with tenets of fairness, suggesting that base rate intrusion is not merely an instantiation of the naturalistic fallacy.
Supplementary Fig. 7. Experiment 5 ($N = 402$). Participants’ average judgments when nonsocial vs. social base rates should not have been used. Each participant was randomly assigned to make two judgments on a 1 to 7 scale (e.g., broken metal spoon vs. broken plastic spoon). All possible pairs of judgments a participant could have been randomly assigned to make are on the x-axes. Statistical tests compare the means of each pair of judgments. Points show the distribution of judgments. Error bars are 95% CIs. *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, n.s. $P > 0.05$. 

![Graph showing judgments for nonsocial and social bases](image-url)
**Supplementary Study**

**Experiment 6.** Experiment 6 was identical to the spoon and gender scenarios in Experiment 2, except sentences that explicitly conveyed disparate outcomes were removed, along with explicit references to base rates. Thus, Experiment 6 remedied only the first and second problems discussed in Experiment 4.

When only these two remedies were implemented, the previous results were again replicated (Supplementary Fig. 8). Violations of the Markov condition strongly depended on whether the base rates were nonsocial or social \([F(1, 393) = 37.65, P < 0.0001]\). In the spoon scenario, which contained nonsocial base rates, the plastic spoon was judged less likely to be kept regardless of whether the two spoons in question were broken \([M_{\text{broken metal}} = 4.81 \text{ vs. } M_{\text{broken plastic}} = 3.89; b = 0.93, t(393) = 5.97, P < 0.0001]\) or intact \([M_{\text{intact metal}} = 6.34 \text{ vs. } M_{\text{intact plastic}} = 5.96; b = 0.38, t(393) = 2.49, P = 0.01]\).

But in the gender scenario, which contained social base rates, judgments erroneously relied upon group differences. A male nurse was judged less likely to be invited than a female nurse \([M_{\text{male nurse}} = 4.76 \text{ vs. } M_{\text{female nurse}} = 5.44; b = -0.67, t(393) = -4.42, P < 0.0001]\), but a female doctor was judged less likely to be invited than a male doctor \([M_{\text{male doctor}} = 6.01 \text{ vs. } M_{\text{female doctor}} = 5.34; b = 0.67, t(393) = 4.37, P < 0.0001]\). As shown previously, social judgments broke with Bayesian rationality and with tenets of fairness, suggesting that base rate intrusion does not occur merely because participants were following Grice’s maxim of relevance. Even when base rates were not explicitly mentioned, they were used, indicating that participants mistakenly brought their own base rate knowledge to bear when making social judgments.
Supplementary Fig. 8. Experiment 6 (N = 397). Participants’ average judgments when nonsocial vs. social base rates should not have been used. Each participant was randomly assigned to make two judgments on a 1 to 7 scale (e.g., broken metal spoon vs. broken plastic spoon). All possible pairs of judgments a participant could have been randomly assigned to make are on the x-axes. Statistical tests compare the means of each pair of judgments. Points show the distribution of judgments. Error bars are 95% CIs. *** P < 0.001, ** P < 0.01, * P < 0.05, n.s. P > 0.05.
Supplementary Study

**Experiment 7.** Experiment 7 was identical to the spoon and gender scenarios in Experiment 2, except sentences that explicitly conveyed disparate outcomes were removed, and the spoons or persons in question were individuated in a manner that made abundantly clear what information was known and what judgments needed to be made. Thus, Experiment 7 remedied only the first and third problems discussed in Experiment 4.

When only these two remedies were implemented, the previous results were again replicated (Supplementary Fig. 9). Violations of the Markov condition strongly depended on whether the base rates were nonsocial or social \([F(1, 399) = 95.12, P < 0.0001]\). In the spoon scenario, which contained nonsocial base rates, the plastic spoon was judged less likely to be kept regardless of whether the two spoons in question were broken \([M_{\text{broken metal}} = 5.33 \text{ vs. } M_{\text{broken plastic}} = 2.94; b = 2.39, t(399) = 14.61, P < 0.0001}\) or intact \([M_{\text{intact metal}} = 6.60 \text{ vs. } M_{\text{intact plastic}} = 5.61; b = 0.99, t(399) = 6.14, P < 0.0001]\).

But in the gender scenario, which contained social base rates, judgments erroneously relied upon group differences. A male nurse was judged less likely to be invited than a female nurse \([M_{\text{male nurse}} = 4.17 \text{ vs. } M_{\text{female nurse}} = 5.04; b = -0.87, t(399) = -5.35, P < 0.0001]\), but a female doctor was judged less likely to be invited than a male doctor \([M_{\text{male doctor}} = 5.76 \text{ vs. } M_{\text{female doctor}} = 4.85; b = 0.91, t(399) = 5.53, P < 0.0001]\). Once more, social judgments broke with Bayesian rationality and with tenets of fairness, indicating that base rate intrusion is not simply an instance of the inverse fallacy.
**Supplementary Fig. 9.** Experiment 7 (N = 403). Participants’ average judgments when nonsocial vs. social base rates should not have been used. Each participant was randomly assigned to make two judgments on a 1 to 7 scale (e.g., broken metal spoon vs. broken plastic spoon). All possible pairs of judgments a participant could have been randomly assigned to make are on the x-axes. Statistical tests compare the means of each pair of judgments. Points show the distribution of judgments. Error bars are 95% CIs. *** P < 0.001, ** P < 0.01, * P < 0.05, n.s. P > 0.05.

---

**Nonsocial**

- More likely
- Less likely

**Social**

- More likely
- Less likely
Supplementary Discussion

Comparing the relative efficacy of strategies to eliminate base rate intrusion. Various strategies were tested to eliminate base rate intrusion. If base rate intrusion does not occur in the gender scenario, then there should be a main effect of profession, no main effect of gender, and no profession by gender interaction. In Supplementary Fig. 10, these effect sizes are plotted from various experiments. It appears that two strategies in particular are effective: removing base rate references and individuating the targets. Visual inspection indicates that the main effect of profession emerges from an additive effect of these two strategies, and that the profession by gender interaction drops close to zero because, again, these two strategies were combined. Across all experiments, no main effect of gender was observed. In Supplementary Fig. 11, the same effect sizes from the analogous nonsocial spoon scenario are plotted. The same strategies that virtually eliminated base rate intrusion in the social domain did not have the same ameliorating effect in the nonsocial domain, as the main effect of material remained and in some experiments even increased.

Experiment 1 is excluded from these comparisons because in this experiment, the relative frequencies of metal vs. plastic spoons or men vs. women were established, which was not the case in the rest of the experiments. Thus, effect size comparisons would be more difficult to interpret. Experiment 3 is also excluded because the within- and between-subjects conditions were switched, again making effect size comparisons more difficult to interpret.

Supplementary Fig. 10. Cross-experimental comparisons of effect sizes in the gender scenario.
Supplementary Fig. 11. Cross-experimental comparisons of effect sizes in the spoon scenario.
Supplementary Discussion

Distribution of judgments in Experiment 2. In Experiment 2, hundreds of participants provided social or nonsocial judgments after reading various scenarios whose content was richly varied. To assess if participants differed in how they made judgments in this experiment, distributions of judgments are shown in contour plots. Distributions of social judgments are shown in Supplementary Fig. 12 and distributions of nonsocial judgments are shown in Supplementary Fig. 13.

Nearly all participants randomly assigned to make social judgments fell prey to base rate intrusion. Very few of these participants’ judgments fall along the 45-degree identity line that indicates equal treatment of two individuals whose social group membership is different.

However, in the nonsocial conditions, some participants were able to properly disregard a spoon’s material, for example. There are clusters of participants whose judgments fall on the 45-degree identity line. These clusters were not present for social judgments, again underscoring differences between the social and nonsocial domains.

Supplementary Fig. 12. Contour plots showing clusters of participants in each of the two pairs of social judgments a participant could have been randomly assigned to make in Experiment 2. The 45-degree identity line indicates accurate judgments that properly set aside social group differences. Red contours indicate higher density of participants; green contours indicate lower density of participants.
Supplementary Fig. 13. Contour plots showing clusters of participants in each of the two pairs of nonsocial judgments a participant could have been randomly assigned to make in Experiment 2. The 45-degree identity line indicates accurate judgments that properly set aside nonsocial differences. **Red** contours indicate higher density of participants; **green** contours indicate lower density of participants.

Collapsed nonsocial judgments akin to two intact spoons of different materials

\[ n = 374 \]

Collapsed nonsocial judgments akin to two broken spoons of different materials

\[ n = 371 \]
Supplementary Methods

Participants, sample size, and informed consent. All participants were recruited from Amazon Mechanical Turk. Sample size and demographic information for each experiment are reported below. These large sample sizes were determined a priori to allow for high power. Analyses for each experiment were conducted after data collection had finished. Harvard University’s Institutional Review Board approved the experiments in this paper. At the beginning of each experiment, participants read and agreed to a consent form.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Sample Size</th>
<th>Demographic Information</th>
</tr>
</thead>
</table>
| 1          | 399         | $M_{age} = 35.75$ years, $SD = 11.36$  
205 females, 192 males, 2 unspecified |
| Replication of Experiment 1 with comprehension check that participants had to answer from memory | 602 | $M_{age} = 35.27$ years, $SD = 11.52$  
349 females, 251 males, 2 unspecified |
| Replication of Experiment 1 with comprehension check that participants could answer by referencing the scenario | 454 | $M_{age} = 34.06$ years, $SD = 11.11$  
247 females, 204 males, 3 unspecified |
| 2          | 1672        | $M_{age} = 33.47$ years, $SD = 10.47$  
779 females, 891 males, 2 unspecified |
| 3          | 398         | $M_{age} = 33.52$, $SD = 11.86$  
187 females, 211 males |
| 4          | 411         | $M_{age} = 33.90$ years, $SD = 11.50$  
218 females, 193 males |
| 5          | 402         | $M_{age} = 35.14$ years, $SD = 11.86$  
192 females, 210 males |
| 6          | 397         | $M_{age} = 34.06$ years, $SD = 11.83$  
193 females, 202 males, 2 unspecified |
| 7          | 403         | $M_{age} = 34.08$ years, $SD = 11.44$  
213 females, 190 males |

Data for Experiment 2 were collected in two rounds. In the first round, the gender and spoon scenarios were replicated. In the second round the remaining scenarios were tested. In the second round, 131 participants were mistakenly recruited who had taken part in the first round. Thus, these participants were excluded.

Data from Experiments 4, 6, and 7 were collected simultaneously. Five participants were excluded for not completing the procedure.
**Experiment 1.** Participants were randomly assigned to either the spoon scenario, which contained nonsocial base rates, or to the gender scenario, which contained social base rates. Both scenarios were discussed in the introduction and their identically structured Bayesian networks are depicted in Fig. 1. After reading the scenario, participants made two judgments about the bottom node given knowledge of the same state of the middle node (e.g., doctor) but different states of the top node (e.g., male vs. female). If base rates are properly set aside, then both judgments should be the same on the 1-7 Likert-type scale (*1 = Extremely unlikely ... 7 = Extremely likely*). All possible sets of judgments that a participant could have been randomly assigned to make, as well as the scenarios themselves, are shown in Supplementary Table 1. To remove any possible memory effects, participants were able to refer to the scenario when making their judgments. The experimenters were aware of which condition participants were randomly assigned to.

**Supplementary Table 1.** Stimuli used in Experiment 1.

<table>
<thead>
<tr>
<th><strong>Nonsocial base rate about spoons</strong></th>
<th><strong>Social base rate about gender</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>There is a factory that makes an equal number of metal spoons and plastic spoons. The factory decides which spoons to keep based solely on whether they are broken or intact. Most of the spoons the factory will keep will be intact so it can sell them. But the factory will also keep some broken spoons to see how they broke. Intact spoons are more likely to be metal than plastic, whereas broken spoons are more likely to be plastic than metal. As a result, the factory will keep more metal spoons than plastic spoons.</td>
<td>There is a charity that has an equal number of men and women it can ask to join an event. The charity asks people to join the event based solely on whether they are a doctor or a nurse. Most of the people the charity will ask will be doctors, who can perform complex procedures. But the charity will also ask some nurses, who can assist the doctors. Doctors are more likely to be male than female, whereas nurses are more likely to be female than male. As a result, the charity will ask more males than females to join the event.</td>
</tr>
<tr>
<td>How likely is the factory to keep each of the following spoons it produced?</td>
<td>How likely is the charity to ask each of the following people to join the event?</td>
</tr>
<tr>
<td><em>1 = Extremely unlikely ... 7 = Extremely likely</em></td>
<td><em>1 = Extremely unlikely ... 7 = Extremely likely</em></td>
</tr>
<tr>
<td>A broken metal spoon</td>
<td>A nurse who is male</td>
</tr>
<tr>
<td>A broken plastic spoon</td>
<td>A nurse who is female</td>
</tr>
<tr>
<td>An intact metal spoon</td>
<td>A doctor who is male</td>
</tr>
<tr>
<td>An intact plastic spoon</td>
<td>A doctor who is female</td>
</tr>
</tbody>
</table>
Analyses. Analyses for all experiments were conducted using R statistical computing’s nlme package. The three-way interaction between base rate (nonsocial vs. social), the middle node in the chain network (broken/nurse vs. intact/doctor), and the top node in the chain network (plastic/female vs. metal/male) was included as a fixed effect. The top node nested within participant was included as a random effect. No other variables were included. In Experiment 3, the between and within-subjects conditions were switched, so the fixed effect remained the same while the random effect was changed to the middle node nested within participant. In Experiment 2, the lme4 package was also used. The fixed effect was again the three-way interaction between base rate type, the middle node, and top node. Random effects for participant and the various scenarios that were tested were also included.

Experiment 2. The procedure was identical to Experiment 1, except the wording of the scenarios was slightly altered and various nonsocial and social base rates were tested. Participants were randomly assigned to read one scenario and then make two judgments, both of which referenced the same state of the middle node but different state of the top node. All stimuli used in Experiment 2, along with the corresponding Bayesian networks, are shown below. The Bayesian networks are presented here for illustrative purposes. They were not presented to participants.
Nonsocial scenarios:

A spoon factory is deciding which spoons to keep based solely on whether they are broken or intact. The factory will keep mostly intact spoons so it can sell them. But the factory will also keep some broken spoons to see how they broke. Intact spoons are more likely to be metal than plastic, whereas broken spoons are more likely to be plastic than metal. As a result, the factory will keep more metal spoons than plastic spoons.

How likely is the factory to keep each of the following spoons it produced?

1 = Extremely unlikely … 7 = Extremely likely

<table>
<thead>
<tr>
<th>Spoons</th>
<th>Corresponding Bayesian network</th>
</tr>
</thead>
<tbody>
<tr>
<td>A broken metal spoon</td>
<td><img src="#" alt="Diagram" /></td>
</tr>
<tr>
<td>A broken plastic spoon</td>
<td></td>
</tr>
<tr>
<td>An intact metal spoon</td>
<td></td>
</tr>
<tr>
<td>An intact plastic spoon</td>
<td></td>
</tr>
</tbody>
</table>
A building's heating system automatically turns on based solely on whether the outside temperature is hot or cold. The heating system will mostly turn on when it is cold outside. But it will sometimes turn on when it is hot outside because of a technical error with the computer system. When the outside temperature is cold, it is more likely to be cloudy than sunny. And when the outside temperature is hot, it is more likely to be sunny than cloudy. As a result, the heating system will turn on more when it is cloudy outside than when it is sunny.

How likely is the heating system to turn on when the weather outside is:

1 = Extremely unlikely ... 7 = Extremely likely

| Hot and cloudy | Cold and cloudy |
| Hot and sunny  | Cold and sunny  |
A restaurant's staff stays late after closing time based solely on whether the restaurant is busy or quiet. The staff will stay late mostly when the restaurant is busy because there is more work to be done. But sometimes, the staff will stay late even if the restaurant is quiet because the manager has important announcements to make. The restaurant is more likely to be busy on weekends than on weekdays, whereas the restaurant is more likely to be quiet on weekdays than on weekends. As a result, the staff is more likely to stay late on weekends than on weekdays.

How likely is the staff to stay late when it is:

1 = Extremely unlikely ... 7 = Extremely likely
A security company calls a home based solely on whether the alarm the goes off or not. The company will call virtually every time the alarm goes off to make sure everything is okay. But sometimes, the company will call even when the alarm hasn’t gone off because of a technical error with the computer system. When the alarm goes off, there is more likely to be a burglary in progress than not. And when the alarm doesn’t go off, there is unlikely to be a burglary in progress. As a result, the company will call mostly when there is a burglary in progress than when there is not.

How likely is the company to call when:

1 = Extremely unlikely ... 7 = Extremely likely

| The alarm isn’t going off and there is a burglary in progress | The alarm is going off and there is a burglary in progress |
| The alarm isn’t going off and there isn’t a burglary in progress | The alarm is going off and there isn’t a burglary in progress |
Social scenarios:

A medical charity is asking people to join an event based solely on whether they are a doctor or a nurse. The charity will ask mostly doctors, who can perform complex procedures. But the charity will also ask some nurses, who can assist the doctors. Doctors are more likely to be male than female, whereas nurses are more likely to be female than male. As a result, the charity will ask more males than females to join the event.

How likely is the charity to ask each of the following people to join the event?

1 = Extremely unlikely ... 7 = Extremely likely

<table>
<thead>
<tr>
<th>A nurse who is male</th>
<th>A doctor who is male</th>
</tr>
</thead>
<tbody>
<tr>
<td>A nurse who is female</td>
<td>A doctor who is female</td>
</tr>
</tbody>
</table>
A news magazine is interviewing people for a story based solely on whether they are an athlete or a scholar. The news magazine will interview mostly scholars since the article is about sports medicine. But the news magazine will also interview some athletes, who bring a different perspective. Scholars are more likely to be White than Black, whereas athletes are more likely to be Black than White. As a result, the news magazine will interview more White people than Black people.

How likely is the news magazine to interview each of the following people?

1 = Extremely unlikely ... 7 = Extremely likely

<table>
<thead>
<tr>
<th>Race (profession)</th>
<th>Corresponding Bayesian network</th>
</tr>
</thead>
<tbody>
<tr>
<td>An athlete who is White</td>
<td>A scholar who is White</td>
</tr>
<tr>
<td>An athlete who is Black</td>
<td>A scholar who is Black</td>
</tr>
</tbody>
</table>
A state prison is interviewing people to get feedback based solely on whether they are a prisoner or a guard. The prison will interview mostly guards since guard safety is a top priority. But the prison will also interview some prisoners, who bring a different perspective. Guards are more likely to be White than Black, whereas prisoners are more likely to be Black than White. As a result, the prison will interview more White people than Black people.

How likely is the prison to interview each of the following people?

1 = Extremely unlikely ... 7 = Extremely likely

| A prisoner who is White | A guard who is White |
| A prisoner who is Black | A guard who is Black |
A radio commercial is casting people based solely on whether they speak with a foreign accent or an American accent. The commercial will cast mostly people with a foreign accent because the commercial is about different languages. But the commercial will also cast some people with an American accent so that there is a contrast. People with a foreign accent are more likely to be foreign citizens than American citizens, whereas people with an American accent are more likely to be American citizens than foreign citizens. As a result, the commercial will cast more foreign citizens than American citizens.

How likely is the commercial to cast each of the following people?

1 = Extremely unlikely ... 7 = Extremely likely

<table>
<thead>
<tr>
<th>Nationality</th>
<th>Corresponding Bayesian network</th>
</tr>
</thead>
<tbody>
<tr>
<td>A person with an American accent who is a foreign citizen</td>
<td>Foreign accent (American accent)</td>
</tr>
<tr>
<td>A person with an American accent who is an American citizen</td>
<td>Foreign citizen (American citizen)</td>
</tr>
<tr>
<td>A person with a foreign accent who is a foreign citizen</td>
<td>Cast (Don’t cast)</td>
</tr>
<tr>
<td>A person with a foreign accent who is an American citizen</td>
<td></td>
</tr>
</tbody>
</table>

176
A medical school is asking people to join a clinical drug trial based solely on whether they are healthy or sick. The medical school will ask mostly sick people to join since the drug needs to be tested. But the medical school will also ask some healthy people to join the drug trial since they are needed for the control condition. Sick people are more likely to be old than young, whereas healthy people are more likely to be young than old. As a result, the medical school will ask more old people than young people to join the clinical drug trial.

How likely is the medical school to ask each of the following people to join the clinical drug trial?

1 = Extremely unlikely ... 7 = Extremely likely

<table>
<thead>
<tr>
<th>A healthy person who is old</th>
<th>A sick person who is old</th>
</tr>
</thead>
<tbody>
<tr>
<td>A healthy person who is young</td>
<td>A sick person who is young</td>
</tr>
</tbody>
</table>
**Experiment 3.** Experiment 3 was identical to the spoon and gender scenarios in Experiment 2, except the judgments participants were randomly assigned to make were different. In the spoon scenario in Experiment 2, judgments were about two spoons of the same status (broken or intact), but of different material (metal vs. plastic). In the gender scenario in Experiment 2, judgments were about two individuals of the same profession (nurse or doctor), but of different gender (male vs. female). In Experiment 3, this was reversed: in the spoon scenario, judgments were about two spoons of the same material, except one was broken and the other was intact. In the gender scenario, judgments were about two persons of the same gender, except one was a nurse and the other was a doctor. The stimuli used are below in Supplementary Table 2.

**Supplementary Table 2.** Stimuli used in Experiment 3.

<table>
<thead>
<tr>
<th>Nonsocial base rate about spoons</th>
<th>Social base rate about gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>A spoon factory is deciding which spoons to keep based solely on whether they are broken or intact. The factory will keep mostly intact spoons so it can sell them. But the factory will also keep some broken spoons to see how they broke. Intact spoons are more likely to be metal than plastic, whereas broken spoons are more likely to be plastic than metal. As a result, the factory will keep more metal spoons than plastic spoons.</td>
<td>A medical charity is asking people to join an event based solely on whether they are a doctor or a nurse. The charity will ask mostly doctors, who can perform complex procedures. But the charity will also ask some nurses, who can assist the doctors. Doctors are more likely to be male than female, whereas nurses are more likely to be female than male. As a result, the charity will ask more males than females to join the event.</td>
</tr>
<tr>
<td>How likely is the factory to keep each of the following spoons it produced?</td>
<td>How likely is the charity to ask each of the following people to join the event?</td>
</tr>
<tr>
<td>1 = Extremely unlikely ... 7 = Extremely likely</td>
<td>1 = Extremely unlikely ... 7 = Extremely likely</td>
</tr>
<tr>
<td>A broken metal spoon</td>
<td>A broken plastic spoon</td>
</tr>
<tr>
<td>An intact metal spoon</td>
<td>An intact plastic spoon</td>
</tr>
<tr>
<td>A nurse who is male</td>
<td>A nurse who is female</td>
</tr>
<tr>
<td>A doctor who is male</td>
<td>A doctor who is female</td>
</tr>
</tbody>
</table>
For clarity, we describe the methods and materials of Experiments 5-7 before providing further details on Experiment 4. Experiment 4 combined all the manipulations in Experiments 5-7.

**Experiment 5.** Experiment 5 was identical to the spoon and gender scenarios in Experiment 2, except the sentences in red font in Supplementary Table 3 were deleted. These sentences explicitly conveyed disparate outcomes about which spoons would be kept or disparate outcomes about which individuals would be invited. To prevent participants from perpetuating these disparate outcomes in their judgments, these sentences were removed. Thus, Experiment 5 remedied only the first problem discussed in Experiment 4.

**Supplementary Table 3.** Stimuli used in Experiment 5.

<table>
<thead>
<tr>
<th>Nonsocial base rate about spoons</th>
<th>Social base rate about gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>A spoon factory is deciding which spoons to keep based solely on whether they are broken or intact. The factory will keep mostly intact spoons so it can sell them. But the factory will also keep some broken spoons to see how they broke. Intact spoons are more likely to be metal than plastic, whereas broken spoons are more likely to be plastic than metal. <strong>As a result, the factory will keep more metal spoons than plastic spoons.</strong> How likely is the factory to keep each of the following spoons it produced? 1 = Extremely unlikely ... 7 = Extremely likely&lt;br&gt;&lt;br&gt;A broken metal spoon&lt;br&gt;A broken plastic spoon&lt;br&gt;An intact metal spoon&lt;br&gt;An intact plastic spoon</td>
<td>A medical charity is asking people to join an event based solely on whether they are a doctor or a nurse. The charity will ask mostly doctors, who can perform complex procedures. But the charity will also ask some nurses, who can assist the doctors. Doctors are more likely to be male than female, whereas nurses are more likely to be female than male. <strong>As a result, the charity will ask more males than females to join the event.</strong> How likely is the charity to ask each of the following people to join the event? 1 = Extremely unlikely ... 7 = Extremely likely&lt;br&gt;&lt;br&gt;A nurse who is male&lt;br&gt;A nurse who is female&lt;br&gt;A doctor who is male&lt;br&gt;A doctor who is female</td>
</tr>
</tbody>
</table>
Experiment 6. Experiment 6 was identical to the spoon and gender scenarios in Experiment 2, except the sentences in red font in Supplementary Table 4 were deleted. As in Experiment 5, the sentences that explicitly conveyed disparate outcomes were also removed. In addition, sentences that explicitly referenced base rates were removed to prevent participants from believing that their judgments should incorporate these base rates. Thus, Experiment 6 remedied only the first and second problems discussed in Experiment 4.

Supplementary Table 4. Stimuli used in Experiment 6.

<table>
<thead>
<tr>
<th>Nonsocial base rate about spoons</th>
<th>Social base rate about gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>A spoon factory is deciding which spoons to keep based solely on whether they are broken or intact. The factory will keep mostly intact spoons so it can sell them. But the factory will also keep some broken spoons to see how they broke. Intact spoons are more likely to be metal than plastic, whereas broken spoons are more likely to be plastic than metal. As a result, the factory will keep more metal spoons than plastic spoons.</td>
<td>A medical charity is asking people to join an event based solely on whether they are a doctor or a nurse. The charity will ask mostly doctors, who can perform complex procedures. But the charity will also ask some nurses, who can assist the doctors. Doctors are more likely to be male than female, whereas nurses are more likely to be female than male. As a result, the charity will ask more males than females to join the event.</td>
</tr>
<tr>
<td>How likely is the factory to keep each of the following spoons it produced?</td>
<td>How likely is the charity to ask each of the following people to join the event?</td>
</tr>
<tr>
<td>1 = Extremely unlikely ... 7 = Extremely likely</td>
<td>1 = Extremely unlikely ... 7 = Extremely likely</td>
</tr>
<tr>
<td>A broken metal spoon</td>
<td>A nurse who is male</td>
</tr>
<tr>
<td>A broken plastic spoon</td>
<td>A nurse who is female</td>
</tr>
<tr>
<td>An intact metal spoon</td>
<td>A doctor who is male</td>
</tr>
<tr>
<td>An intact plastic spoon</td>
<td>A doctor who is female</td>
</tr>
</tbody>
</table>
**Experiment 7.** Experiment 7 was identical to the spoon and gender scenarios in Experiment 2, except the sentences in red font in Supplementary Table 5 were deleted. As in Experiment 5 and 6, the sentences that explicitly conveyed disparate outcomes were removed. The questions were also altered so that the spoons or persons in question were individuated in a manner that made abundantly clear what information was known and what judgments needed to be made. These changes are marked in green font, whereas the previous question wording is in red font. Thus, Experiment 7 remedied only the first and third problems discussed in Experiment 4. Explicit references to base rates still remained.
Supplementary Table 5. Stimuli used in Experiment 7.

<table>
<thead>
<tr>
<th>Nonsocial base rate about spoons</th>
<th>Social base rate about gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>A spoon factory is deciding which spoons to keep based solely on whether they are broken or intact. The factory will keep mostly intact spoons so it can sell them. But the factory will also keep some broken spoons to see how they broke. Intact spoons are more likely to be metal than plastic, whereas broken spoons are more likely to be plastic than metal. As a result, the factory will keep more metal spoons than plastic spoons.</td>
<td>A medical charity is asking people to join an event based solely on whether they are a doctor or a nurse. The charity will ask mostly doctors, who can perform complex procedures. But the charity will also ask some nurses, who can assist the doctors. Doctors are more likely to be male than female, whereas nurses are more likely to be female than male. As a result, the charity will ask more males than females to join the event.</td>
</tr>
<tr>
<td><strong>How likely is the factory to keep each of the following spoons it produced?</strong></td>
<td><strong>How likely is the charity to ask each of the following people to join the event?</strong></td>
</tr>
<tr>
<td>1 = Extremely unlikely ... 7 = Extremely likely</td>
<td>1 = Extremely unlikely ... 7 = Extremely likely</td>
</tr>
<tr>
<td>A broken metal spoon</td>
<td>A nurse who is male</td>
</tr>
<tr>
<td>A broken plastic spoon</td>
<td>A doctor who is male</td>
</tr>
<tr>
<td>1 = Extremely unlikely ... 7 = Extremely likely</td>
<td>1 = Extremely unlikely ... 7 = Extremely likely</td>
</tr>
<tr>
<td>The factory produced a spoon whose unique serial number is WUL5a. This spoon is made of metal and it is broken. The factory hasn’t decided yet of it will keep it. How likely is the factory to keep this spoon?</td>
<td>The charity has a file on a person whose unique identifier is WUL5a. This person is a male nurse. The charity hasn’t decided yet if it will ask this person to join the event. How likely is the charity to ask this person?</td>
</tr>
<tr>
<td>The factory produced a spoon whose unique serial number is AeT2k. This spoon is made of plastic and it is broken. The factory hasn’t decided yet of it will keep it. How likely is the factory to keep this spoon?</td>
<td>The charity has a file on a person whose unique identifier is AeT2k. This person is a female nurse. The charity hasn’t decided yet if it will ask this person to join the event. How likely is the charity to ask this person?</td>
</tr>
<tr>
<td>The factory produced a spoon whose unique serial number is WUL5a. This spoon is made of metal and it is intact. The factory hasn’t decided yet of it will keep it. How likely is the factory to keep this spoon?</td>
<td>The charity has a file on a person whose unique identifier is WUL5a. This person is a male doctor. The charity hasn’t decided yet if it will ask this person to join the event. How likely is the charity to ask this person?</td>
</tr>
<tr>
<td>The factory produced a spoon whose unique serial number is AeT2k. This spoon is made of plastic and it is intact. The factory hasn’t decided yet of it will keep it. How likely is the factory to keep this spoon?</td>
<td>The charity has a file on a person whose unique identifier is AeT2k. This person is a female doctor. The charity hasn’t decided yet if it will ask this person to join the event. How likely is the charity to ask this person?</td>
</tr>
</tbody>
</table>
**Experiment 4.** Experiment 4 was identical to the spoon and gender scenarios in Experiment 2, except the sentences in red font in Supplementary Table 6 were deleted. As before, the sentences that explicitly conveyed disparate outcomes were removed. Furthermore, sentences that explicitly referenced the base rates were also removed. And finally, the questions were also changed so that the spoons or persons in question were individuated in a manner that made abundantly clear what information was known and what judgment needed to be made. These changes are marked in green font, whereas the previous question wording is in red font. Thus, all three of the problems discussed in Experiment 4 were remedied.
Supplementary Table 6. Stimuli used in Experiment 4.

<table>
<thead>
<tr>
<th>Nonsocial base rate about spoons</th>
<th>Social base rate about gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>A spoon factory is deciding which spoons to keep based solely on whether they are broken or intact. The factory will keep mostly intact spoons so it can sell them. But the factory will also keep some broken spoons to see how they broke. Intact spoons are more likely to be metal than plastic, whereas broken spoons are more likely to be plastic than metal. As a result, the factory will keep more metal spoons than plastic spoons.</td>
<td>A medical charity is asking people to join an event based solely on whether they are a doctor or a nurse. The charity will ask mostly doctors, who can perform complex procedures. But the charity will also ask some nurses, who can assist the doctors. Doctors are more likely to be male than female, whereas nurses are more likely to be female than male. As a result, the charity will ask more males than females to join the event.</td>
</tr>
</tbody>
</table>

How likely is the factory to keep each of the following spoons it produced?

<table>
<thead>
<tr>
<th>1 = Extremely unlikely ... 7 = Extremely likely</th>
</tr>
</thead>
<tbody>
<tr>
<td>A broken metal spoon</td>
</tr>
<tr>
<td>A broken plastic spoon</td>
</tr>
<tr>
<td>An intact metal spoon</td>
</tr>
<tr>
<td>An intact plastic spoon</td>
</tr>
</tbody>
</table>

1 = Extremely unlikely ... 7 = Extremely likely

The factory produced a spoon whose unique serial number is WUL5a. This spoon is made of metal and it is broken. The factory hasn’t decided yet of it will keep it. How likely is the factory to keep this spoon?

The factory produced a spoon whose unique serial number is WUL5a. This spoon is made of metal and it is intact. The factory hasn’t decided yet of it will keep it. How likely is the factory to keep this spoon?

The charity has a file on a person whose unique identifier is WUL5a. This person is a male nurse. The charity hasn’t decided yet if it will ask this person to join the event. How likely is the charity to ask this person?

The charity has a file on a person whose unique identifier is WUL5a. This person is a male doctor. The charity hasn’t decided yet if it will ask this person to join the event. How likely is the charity to ask this person?

The factory produced a spoon whose unique serial number is AeT2k. This spoon is made of plastic and it is broken. The factory hasn’t decided yet of it will keep it. How likely is the factory to keep this spoon?

The factory produced a spoon whose unique serial number is AeT2k. This spoon is made of plastic and it is intact. The factory hasn’t decided yet of it will keep it. How likely is the factory to keep this spoon?

The charity has a file on a person whose unique identifier is AeT2k. This person is a female nurse. The charity hasn’t decided yet if it will ask this person to join the event. How likely is the charity to ask this person?

The charity has a file on a person whose unique identifier is AeT2k. This person is a female doctor. The charity hasn’t decided yet if it will ask this person to join the event. How likely is the charity to ask this person?
**Supplementary Table 7.** Sample size, mean, and standard deviation in each cell of all experiments.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Condition</th>
<th>n</th>
<th>Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Nonsocial – spoons: broken metal spoon vs. broken plastic spoon</td>
<td>102</td>
<td>broken metal spoon: 4.91 (1.85) broken plastic spoon: 3.73 (1.73)</td>
</tr>
<tr>
<td>1</td>
<td>Nonsocial – spoons: intact metal spoon vs. intact metal spoon</td>
<td>100</td>
<td>intact metal spoon: 6.52 (1.00) intact plastic spoon: 5.10 (1.97)</td>
</tr>
<tr>
<td>1</td>
<td>Social – gender: male nurse vs. female nurse</td>
<td>101</td>
<td>male nurse: 3.70 (1.53) female nurse: 5.58 (1.38)</td>
</tr>
<tr>
<td>1</td>
<td>Social – gender: male doctor vs. female doctor</td>
<td>96</td>
<td>male doctor: 6.18 (0.93) female doctor: 4.36 (1.81)</td>
</tr>
<tr>
<td>Replication of Exp. 1 with comprehension check that participants had to answer from memory</td>
<td>Nonsocial – spoons: broken metal spoon vs. broken plastic spoon</td>
<td>154</td>
<td>For only participants who passed comprehension check: broken metal spoon: 4.85 (1.78) broken plastic spoon: 3.73 (1.60)</td>
</tr>
<tr>
<td>Replication of Exp. 1 with comprehension check that participants had to answer from memory</td>
<td>Nonsocial – spoons: intact metal spoon vs. intact metal spoon</td>
<td>144</td>
<td>For only participants who passed comprehension check: intact metal spoon: 6.76 (0.64) intact plastic spoon: 6.08 (1.55)</td>
</tr>
<tr>
<td>Replication of Exp. 1 with comprehension check that participants had to answer from memory</td>
<td>Social – gender: male nurse vs. female nurse</td>
<td>154</td>
<td>For only participants who passed comprehension check: male nurse: 3.71 (1.65) female nurse: 5.42 (1.60)</td>
</tr>
<tr>
<td>Replication of Exp. 1 with comprehension check that participants had to answer from memory</td>
<td>Social – gender: male doctor vs. female doctor</td>
<td>150</td>
<td>For only participants who passed comprehension check: male doctor: 6.29 (0.98) female doctor: 4.77 (1.75)</td>
</tr>
</tbody>
</table>
**Supplementary Table 7 continued.** Sample size, mean, and standard deviation in each cell of all experiments.

<table>
<thead>
<tr>
<th>Replication of Exp. 1 with comprehension check that participants could answer by referencing the scenario</th>
<th>Nonsocial – spoons: broken metal spoon vs. broken plastic spoon</th>
<th>112</th>
<th>For only participants who passed comprehension check: broken metal spoon: 4.35 (1.77) broken plastic spoon: 3.69 (1.67)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replication of Exp. 1 with comprehension check that participants could answer by referencing the scenario</td>
<td>Nonsocial – spoons: intact metal spoon vs. intact metal spoon</td>
<td>113</td>
<td>For only participants who passed comprehension check: intact metal spoon: 6.69 (0.88) intact plastic spoon: 5.78 (1.63)</td>
</tr>
<tr>
<td>Replication of Exp. 1 with comprehension check that participants could answer by referencing the scenario</td>
<td>Social – gender: male nurse vs. female nurse</td>
<td>115</td>
<td>For only participants who passed comprehension check: male nurse: 3.61 (1.57) female nurse: 5.68 (1.06)</td>
</tr>
<tr>
<td>Replication of Exp. 1 with comprehension check that participants could answer by referencing the scenario</td>
<td>Social – gender: male doctor vs. female doctor</td>
<td>114</td>
<td>For only participants who passed comprehension check: male doctor: 6.22 (0.94) female doctor: 4.29 (1.73)</td>
</tr>
<tr>
<td>2</td>
<td>Nonsocial – spoons: broken metal spoon vs. broken plastic spoon</td>
<td>98</td>
<td>broken metal spoon: 4.56 (1.83) broken plastic spoon: 4.09 (1.63)</td>
</tr>
<tr>
<td>2</td>
<td>Nonsocial – spoons: intact metal spoon vs. intact metal spoon</td>
<td>99</td>
<td>intact metal spoon: 6.46 (0.91) intact plastic spoon: 4.93 (1.85)</td>
</tr>
<tr>
<td>2</td>
<td>Nonsocial – weather: hot cloudy vs. hot sunny</td>
<td>85</td>
<td>hot cloudy: 4.38 (1.60) hot sunny: 2.74 (1.89)</td>
</tr>
<tr>
<td>2</td>
<td>Nonsocial – weather: cold cloudy vs. cold sunny</td>
<td>88</td>
<td>cold cloudy: 6.45 (0.76) cold sunny: 4.68 (1.46)</td>
</tr>
<tr>
<td>2</td>
<td>Nonsocial – days of the week: quiet weekend vs. quiet weekday</td>
<td>93</td>
<td>quiet weekend: 4.32 (1.50) quiet weekday: 3.29 (1.32)</td>
</tr>
<tr>
<td>2</td>
<td>Nonsocial – days of the week: busy weekend vs. busy weekday</td>
<td>93</td>
<td>busy weekend: 6.41 (0.96) busy weekday: 5.40 (1.39)</td>
</tr>
<tr>
<td>2</td>
<td>Nonsocial – alarms: no alarm, burglary vs. no alarm, no</td>
<td>95</td>
<td>no alarm, burglary: 2.54 (1.62) no alarm, no burglary: 3.48 (1.87)</td>
</tr>
</tbody>
</table>
Supplementary Table 7 continued. Sample size, mean, and standard deviation in each cell of all experiments.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Condition</th>
<th>n</th>
<th>Mean (SD)</th>
</tr>
</thead>
</table>
| 2          | Social – gender: male nurse vs. female nurse | 101 | male nurse: 3.74 (1.57)  
                          |           |                  | female nurse: 5.62 (1.20) |
| 2          | Social – gender: male doctor vs. female doctor | 101 | male doctor: 6.15 (0.92)  
                          |           |                  | female doctor: 4.12 (1.56) |
| 2          | Social – race (profession) white athlete vs. black athlete | 96 | white athlete: 3.69 (1.46)  
                          |           |                  | black athlete: 5.54 (1.27) |
| 2          | Social – race (profession) white scholar vs. black scholar | 94 | white scholar: 6.14 (0.87)  
                          |           |                  | black scholar: 3.48 (1.28) |
| 2          | Social – race (crime) white prisoner vs. black prisoner | 87 | white prisoner: 3.83 (1.61)  
                          |           |                  | black prisoner: 5.33 (1.45) |
| 2          | Social – race (crime) white guard vs. black guard | 91 | white guard: 6.12 (0.93)  
                          |           |                  | black guard: 3.18 (1.19) |
| 2          | Social – nationality American accent, foreigner vs. American accent, American | 86 | American accent, foreigner: 3.57 (1.67)  
                          |           |                  | American accent, American: 5.24 (1.48) |
| 2          | Social – nationality foreign accent, foreigner vs. foreign accent, American | 95 | foreign accent, foreigner: 5.66 (1.29)  
                          |           |                  | foreign accent, American: 3.65 (1.46) |
| 2          | Social – age healthy old vs. healthy young | 89 | healthy old: 4.02 (1.69)  
                          |           |                  | healthy young: 4.54 (1.64) |
| 2          | Social – age sick old vs. sick young | 87 | sick old: 6.28 (1.06)  
                          |           |                  | sick young: 4.11 (1.46) |
| 3          | Nonsocial – spoons: broken metal spoon vs. intact metal spoon | 93 | broken metal spoon: 4.13 (1.71)  
                          |           |                  | intact metal spoon: 6.54 (1.07) |
| 3          | Nonsocial – spoons: broken plastic spoon vs. intact plastic spoon | 97 | broken plastic spoon: 3.82 (1.59)  
                          |           |                  | intact plastic spoon: 6.07 (1.38) |
| 3          | Social – gender: male nurse vs. male doctor | 105 | male nurse: 3.55 (1.61)  
                          |           |                  | male doctor: 6.25 (1.06) |
| 3          | Social – gender: female nurse vs. female doctor | 103 | female nurse: 5.81 (1.11)  
                          |           |                  | female doctor: 3.82 (1.62) |
| 4          | Nonsocial – spoons: broken metal spoon vs. broken plastic spoon | 105 | broken metal spoon: 4.70 (1.50)  
                          |           |                  | broken plastic spoon: 3.13 (1.55) |
| 4          | Nonsocial – spoons: intact metal spoon vs. intact metal spoon | 102 | intact metal spoon: 6.38 (1.29)  
                          |           |                  | intact plastic spoon: 6.13 (1.27) |
| 4          | Social – gender: male nurse vs. female nurse | 102 | male nurse: 4.46 (1.18)  
                          |           |                  | female nurse: 4.61 (1.25) |
### Supplementary Table 7 continued

Sample size, mean, and standard deviation in each cell of all experiments.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Condition</th>
<th>n</th>
<th>Mean (SD)</th>
</tr>
</thead>
</table>
| 4          | Social – gender: male doctor vs. female doctor | 102 | male doctor: 5.76 (0.96)  
female doctor: 5.51 (1.12) |
| 5          | Nonsocial – spoons: broken metal spoon vs. broken plastic spoon | 102 | broken metal spoon: 4.69 (1.91)  
broken plastic spoon: 3.57 (1.63) |
| 5          | Nonsocial – spoons: intact metal spoon vs. intact metal spoon | 101 | intact metal spoon: 6.44 (1.20)  
intact plastic spoon: 5.49 (1.92) |
| 5          | Social – gender: male nurse vs. female nurse | 96  | male nurse: 3.95 (1.60)  
female nurse: 5.71 (1.11) |
| 5          | Social – gender: male doctor vs. female doctor | 103 | male doctor: 6.16 (0.99)  
female doctor: 4.78 (1.60) |
| 6          | Nonsocial – spoons: broken metal spoon vs. broken plastic spoon | 97  | broken metal spoon: 4.81 (1.52)  
broken plastic spoon: 3.89 (1.61) |
| 6          | Nonsocial – spoons: intact metal spoon vs. intact metal spoon | 99  | intact metal spoon: 6.34 (1.10)  
intact plastic spoon: 5.96 (1.35) |
| 6          | Social – gender: male nurse vs. female nurse | 101 | male nurse: 4.76 (1.30)  
female nurse: 5.44 (1.28) |
| 6          | Social – gender: male doctor vs. female doctor | 100 | male doctor: 6.01 (1.07)  
female doctor: 5.34 (1.42) |
| 7          | Nonsocial – spoons: broken metal spoon vs. broken plastic spoon | 100 | broken metal spoon: 5.33 (1.46)  
broken plastic spoon: 2.94 (1.48) |
| 7          | Nonsocial – spoons: intact metal spoon vs. intact metal spoon | 103 | intact metal spoon: 6.60 (0.62)  
intact plastic spoon: 5.61 (1.48) |
| 7          | Social – gender: male nurse vs. female nurse | 101 | male nurse: 4.17 (1.41)  
female nurse: 5.04 (1.27) |
| 7          | Social – gender: male doctor vs. female doctor | 99  | male doctor: 5.76 (0.89)  
female doctor: 4.85 (1.47) |