The base rate principle and the fairness principle in social judgment

Jack Cao[a,1] and Mahzarin R. Banaji[a]

[a]Department of Psychology, Harvard University, Cambridge, MA 02138

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved May 17, 2016 (received for review December 15, 2015)

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. However, 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. Taken together, these experiments demonstrate that explicit beliefs uphold fairness and incorporate obvious and relevant facts, but implicit beliefs uphold base rates and appear relatively impervious to counterstereotypic facts.

social cognition | stereotypes | base rates | fairness | Implicit Association Test

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 (1) and people with violent felonies on their record are more likely to be Black than White (2). In fact, because group membership contains useful information for deciding whether an individual has a certain attribute, stereotypes have been conceptualized as base rates (3–6). Moreover, decision theorists have shown that base rates are critical ingredients for making predictions (7, 8), as neglecting base rates will cause predictions to deviate from what is statistically likely (9).

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 (10, 11), and the European Union has banned gender-based insurance premiums (12).

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 Exp. 1, participants meet Jonathan and Elizabeth and must predict who is the doctor and who is the nurse. If participants use base rates, then Jonathan will be more likely than Elizabeth to be the doctor. However, if participants privilege fairness, then both Jonathan and Elizabeth will be equally likely to be the doctor.* Next, participants were taught one of three types of individuating facts: (*i*) stereotypic facts: Jonathan is the doctor and Elizabeth is the nurse; (*ii*) counterstereotypic facts: Elizabeth is the doctor and Jonathan is the nurse; or (*iii*) irrelevant facts that served as a control: Jonathan vacationed in Colorado and 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. However, 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 (13). 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.

Significance

In the absence of individuating facts, beliefs about individuals can either use 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. Whereas 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.

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

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: All data are available on the Open Science Framework (osf.io/gaedk).

[1]To whom correspondence should be addressed. Email: jackcao@fas.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1524268113/-/DCSupplemental.

*We do not imply that the base rate and fairness principles are mutually exclusive. Our measures allow participants to use both principles in their initial beliefs.

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) (14), 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 (15, 16), even though they can be dissociated (17).

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 afterward 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. However, 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 (18, 19), making it seem unlikely that implicit beliefs will incorporate individuating facts like explicit beliefs would. However, other research has identified conditions under which implicit associations appear highly amenable to new information (20–24).[†] 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 (24), 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 with 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 (18, 22, 27). 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. However, 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. However, 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 (28). 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?[‡]

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 [mean (M) = −0.30, one-sample $t(573) = -11.74$, $P < 0.0001$, 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 $t(573) = -27.23$, $P < 0.0001$, Cohen's $d = 1.14$].

Both explicit and implicit beliefs were, on average, consistent with base rate use. 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. $M_{before} = -0.24$; $P = 0.19$). In the experimental conditions, explicit beliefs were correctly updated to align with both stereotypic facts ($M_{after} = -2.42$ vs. $M_{before} = -0.34$; $P < 0.0001$) and counterstereotypic facts ($M_{after} = 2.64$ vs. $M_{before} = -0.32$; $P < 0.0001$). Thus, when individuating facts made absolutely clear each person's profession, explicit beliefs appropriately set aside the fairness principle (Fig. S2).

Although explicit beliefs displayed a sensible pattern of updating, implicit beliefs aligned only with stereotypic facts. When individuating facts were counterstereotypic, implicit beliefs still used 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 such that implicit beliefs were closer to a neutral *D* score of zero ($M_{after} = -0.30$ vs. $M_{before} = -0.43$; $P < 0.0001$).[§] Stereotypic facts had an effect above and beyond test–retest in the expected direction [$t(571) = -3.68$, $P = 0.0003$], resulting in *D*

[†]We use the term "implicit belief" because implicit associations between individual and profession are truth evaluable when compared with 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 (25, 26). The use of the term implicit belief should not be interpreted as a position on this debate, which this paper does not address.

[‡]Some readers may note the relevance of Bayesian models. Although 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 (3). 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.

[§]By taking the IAT twice, participants improve their ability sort stimuli quickly to the appropriate category, leading to reduced effects (29).

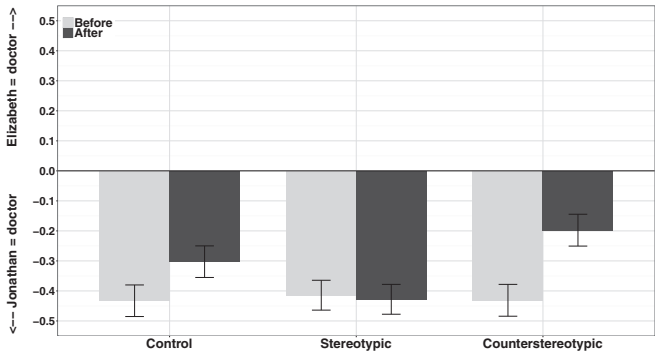

Fig. 1. Exp. 1 (N = 574): Mean implicit beliefs about Jonathan and Elizabeth's professions. IAT D Scores are on the y axis. Error bars are 95% CIs.

scores consistent with each person's actual profession [M_{after} = −0.43, 95% confidence interval (CI) (−0.48, −0.38)].

Counterstereotypic facts also had an effect above and beyond test–retest in the expected direction [$t(571) = 2.57$, $P = 0.01$]. 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, whereas Elizabeth, who is actually the doctor, was more strongly associated with nurse.[¶] Despite the certainty provided by these individuating facts, implicit beliefs continued to rely on the base rate principle.

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.

However, 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 [$t(78) = -0.13$, $P = 0.90$]. 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.

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 Exp. 1 with a new set of names and professions. In Exp. 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 $t(807) = -8.98$, $P < 0.0001$, 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 $t(807) = -23.20$, $P < 0.0001$, Cohen's d = 0.82].

As before, visual inspection of each belief distribution reveals alignment with a different principle (Fig. S3). Although 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. $M_{before} = -0.13$; $P < 0.0001$). Upon learning counterstereotypic facts that Jennifer is the scientist and Richard is the artist, explicit beliefs reflected these facts ($M_{after} = 2.53$ vs. $M_{before} = -0.21$; $P < 0.0001$).

However, as in Exp. 1, implicit beliefs aligned only with stereotypic facts. After learning counterstereotypic facts, implicit beliefs about Richard and Jennifer's professions still used 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 [$t(805) = -2.98$, $P = 0.003$], resulting in D scores consistent with Richard's actual profession as the scientist and Jennifer's actual profession as the artist [M_{after} = −0.29, 95% CI (−0.33, −0.25)].

Counterstereotypic facts also had an effect above and beyond test–retest in the expected direction [$t(805) = 4.03$, $P = 0.0001$]. However, this effect was insufficient to align implicit beliefs with the fact that Jennifer, not Richard, is the scientist [M_{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_{after} = −0.18, 95% CI (−0.24, −0.12)]. However, among these participants who learned counterstereotypic facts, implicit beliefs once again failed to incorporate these facts [M_{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 [$ts(169) < |1.56|$, $Ps > 0.12$]. 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

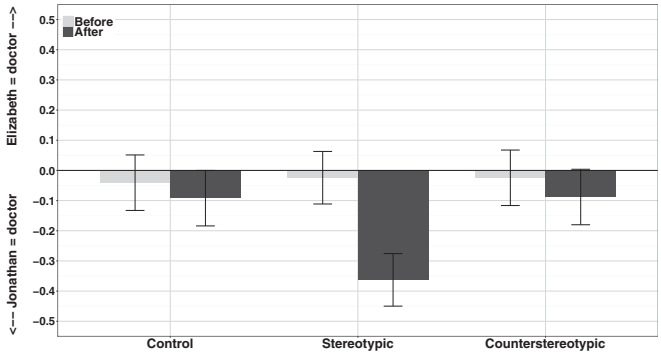

Fig. 2. Exp. 1: Mean implicit beliefs about Jonathan and Elizabeth's professions only among participants whose implicit beliefs before learning the facts were neutral (N = 81). IAT D scores are on the y axis. Error bars are 95% CIs.

[¶]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. (27) used a Solomon four-group design (i.e., random assignment to pretest) 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.

specific individuals but rather as generic representatives of the many individuals who share these familiar names. Exp. 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 Exps. 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 Exp. 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, one-sample $t(658) = -5.88$, $P < 0.0001$, 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, one-sample $t(658) = -21.83$, $P < 0.0001$, 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 ($M_{after} = -2.62$ vs. $M_{before} = -0.07$; $P < 0.0001$) or counterstereotypic ($M_{after} = 2.50$ vs. $M_{before} = -0.20$; $P < 0.0001$).

In contrast, implicit beliefs about Lapper and Affina's professions aligned only with stereotypic facts, thereby replicating the previous results, but with novel names. After learning that 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 [$t(656) = -2.38$, $P = 0.018$], 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)]. Although counterstereotypic facts also had an effect above and beyond test–retest [$t(656) = 2.03$, $P = 0.04$], 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 and 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)], although the effect of these facts did not exceed test–retest [$t(140) = -1.33$, $P = 0.19$]. 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 [$t(140) = 2.02$, $P = 0.046$]. 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. 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 ($N = 1{,}417$) reported explicit beliefs that slightly favored Matthew to be the scientist and Benjamin to be the artist [M = −0.03, one-sample $t(1{,}416) = -2.13$, $P = 0.03$, 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, one-sample $t(1{,}416) = 3.03$, $P = 0.003$, Cohen's $d = 0.08$].

The large sample size magnifies this baseline difference from zero, which is negligible (Cohen's ds < 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 ($M_{after} = -2.50$ vs. $M_{before} = -0.05$; $P < 0.0001$) or vice versa ($M_{after} = 2.16$ vs. $M_{before} = 0.03$; $P < 0.0001$).

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 [$M_{after} = -0.06$, 95% CI (−0.10, −0.03)]. The effect of these individuating facts exceeded test–retest [$t(1{,}414) = -4.67$, $P < 0.0001$]. Additionally, when Benjamin was the scientist and Matthew was the artist, implicit beliefs aligned with these facts [$M_{after} = 0.10$, 95% CI (0.06, 0.14)]. The effect of these individuating facts came extremely close to significantly exceeding test-retest [$t(1{,}414) = 1.95$, $P = 0.052$].

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 [$M_{after} = -0.10$, 95% CI (−0.14, −0.05)], an effect that exceeded test–retest [$t(400) = -2.61$, $P = 0.009$]. 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.

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 but 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,

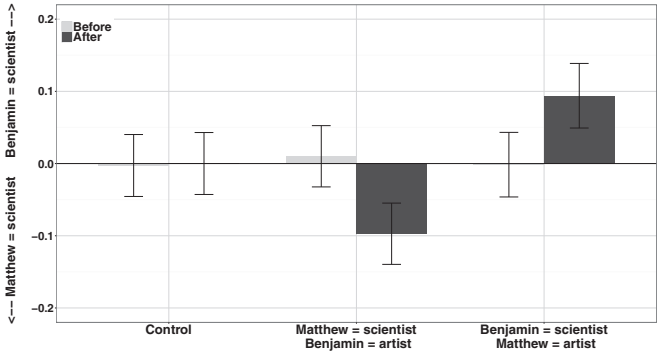

Fig. 3. Exp. 4: Mean implicit beliefs about Matthew and Benjamin's professions only among participants whose implicit beliefs before learning the facts were neutral ($N = 403$). IAT D scores are on the y axis. Error bars are 95% CIs.

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, whereas past work has focused on dissociations between explicit and implicit responses, we assess how both responses compare against two kinds of real-world facts: (*i*) statistical regularities that can be applied and (*ii*) 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 ‡, 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 (30, 31). 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 toward judgments of entire groups, then a single counterstereotypic example need not lead to a dramatic shift in belief. However, 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 (*i*) work in domains that are deeply social in nature where the base rate principle and fairness principle can be in conflict, and (*ii*) use implicit measures alongside standard explicit measures to test the boundary conditions of people's ability to update their beliefs. Much of cognition occurs unconsciously (32) and the topic of changing implicit responses has gained traction (18–24, 27). 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 and Ferguson's (24) finding that highly diagnostic information can lead to substantial revisions of implicit evaluations. However, these researchers sought to update implicit evaluations of nonstereotyped individuals, whereas we sought to update implicit beliefs about stereotyped individuals. Moreover, Cone and 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 (33, 34). 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, and Zingales (35), 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. Exp. 3, which used unfamiliar names and replicated the findings of Exps. 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 (36). 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 Exp. 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. 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 Materials and Methods* 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 Materials and Methods* for all experimental stimuli.

IAT Scoring Procedure. Following Greenwald, Nosek, and Banaji (37), 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 (38). 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, and 4. These experiments were identical to Exp. 1 except the names and professions were changed accordingly. In Exp. 2, participants answered four questions that tested knowledge of Lapper and Affina's gender.

ACKNOWLEDGMENTS. We thank Melissa Ferguson and Lee Jussim for helpful comments on earlier drafts. We also thank Christopher Willis and Natalia Dashan for assistance with stimuli construction. This work was supported by a National Science Foundation Graduate Research Fellowship DGE 1144152 (to J.C.).

1. Kaiser Family Foundation (2015) *Distribution of Physicians by Gender*. Available at kff.org/other/state-indicator/physicians-by-gender. Accessed December 15, 2015.
2. Bureau of Justice Statistics (2015) *Violent Felons in Large Urban Counties*. Available at www.bjs.gov/content/pub/pdf/vfluc.pdf. Accessed December 15, 2015.
3. Locksley A, Borgida E, Brekke N, Hepburn C (1980) Sex stereotypes and social judgment. *J Pers Soc Psychol* 39(5):821–831.
4. Locksley A, Hepburn C, Ortiz V (1982) Social stereotypes and judgments of individuals: An instance of the base-rate fallacy. *J Exp Soc Psychol* 18(1):23–42.
5. Krosnick JA, Li F, Lehman DR (1990) Conversational conventions, order of information acquisition, and the effect of base rates on individuating information on social judgments. *J Pers Soc Psychol* 59(6):1140–1152.
6. Jussim L (2012) *Social Perception and Social Reality: Why Accuracy Dominates Bias and Self-Fulfilling Prophecy* (Oxford Univ Press, New York).
7. Beyth-Marom R, Fischhoff B (1983) Diagnosticity and pseudodiagnosticity. *J Pers Soc Psychol* 45(6):1185–1195.
8. Bar-Hillel M (1980) The base-rate fallacy in probability judgments. *Acta Psychol (Amst)* 44(3):211–233.
9. Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185(4157):1124–1131.
10. Koehler JJ (1992) Probabilities in the courtroom: An evaluation of the objections and policies. *Handbook of Psychology and Law*, eds Kagehiro DK, Laufer WS (Springer, New York), pp 167–184.
11. Schauer F (2003) *Profiles, Probabilities, and Stereotypes* (Belknap Press of Harvard Univ Press, Cambridge, MA).
12. Test-Achats v. Council of Ministers. *European Court of Justice* (2011).
13. Fiske ST, Neuberg SL (1990) A continuum of impression formation, from category based to individuating processes: Influences of information and motivation on attention and interpretation. *Adv Exp Soc Psychol* 23:1–74.
14. Greenwald AG, McGhee DE, Schwartz JL (1998) Measuring individual differences in implicit cognition: The implicit association test. *J Pers Soc Psychol* 74(6):1464–1480.
15. Cameron CD, Brown-Iannuzzi JL, Payne BK (2012) Sequential priming measures of implicit social cognition: A meta-analysis of associations with behavior and explicit attitudes. *Pers Soc Psychol Rev* 16(4):330–350.
16. Greenwald AG, Poehlman TA, Uhlmann EL, Banaji MR (2009) Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *J Pers Soc Psychol* 97(1):17–41.
17. Nosek BA, et al. (2007) Pervasiveness and correlates of implicit attitudes and stereotypes. *Eur Rev Soc Psychol* 18(1):36–88.
18. Gregg AP, Seibt B, Banaji MR (2006) Easier done than undone: asymmetry in the malleability of implicit preferences. *J Pers Soc Psychol* 90(1):1–20.
19. Rydell RJ, McConnell AR, Strain LM, Claypool HM, Hugenbeg K (2007) Implicit and explicit attitudes respond differently to increasing amounts of counterattitudinal information. *Eur J Soc Psychol* 37(5):867–878.
20. Blair IV (2002) The malleability of automatic stereotypes and prejudice. *Pers Soc Psychol Rev* 6(3):242–261.
21. Blair IV, Ma JE, Lenton AP (2001) Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *J Pers Soc Psychol* 81(5):828–841.
22. Wyer NA (2016) Easier done than undone…by some of the people, some of the time: The role of elaboration in explicit and implicit group preferences. *J Exp Soc Psychol* 63:77–85.
23. Wyer NA (2010) You never get a second change to make a first (implicit) impression: The role of elaboration in the formation and revision of implicit impressions. *Soc Cogn* 28(1):1–19.
24. Cone J, Ferguson MJ (2015) He did what? The role of diagnosticity in revising implicit evaluations. *J Pers Soc Psychol* 108(1):37–57.
25. Gawronski B, Bodenhausen GV (2006) Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychol Bull* 132(5):692–731.
26. Mitchell CJ, De Houwer J, Lovibond PF (2009) The propositional nature of human associative learning. *Behav Brain Sci* 32(2):183–198, discussion 198–246.
27. Lai CK, et al. (2014) Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *J Exp Psychol Gen* 143(4):1765–1785.
28. Heiphetz L, Spelke ES, Harris PL, Banaji MR (2013) The development of reasoning about beliefs: Fact, preference, and ideology. *J Exp Soc Psychol* 49(3):559–565.
29. Cunningham WA, Preacher KJ, Banaji MR (2001) Implicit attitude measures: consistency, stability, and convergent validity. *Psychol Sci* 12(2):163–170.
30. Griffiths TL, Tenenbaum JB (2006) Optimal predictions in everyday cognition. *Psychol Sci* 17(9):767–773.
31. Xu F, Tenenbaum JB (2007) Word learning as Bayesian inference. *Psychol Rev* 114(2): 245–272.
32. Hassin RR, Uleman JS, Bargh JA (2005) *The New Unconscious* (Oxford Univ Press, New York).
33. Sharot T, et al. (2012) Selectively altering belief formation in the human brain. *Proc Natl Acad Sci USA* 109(42):17058–17062.
34. Eil D, Rao JM (2011) The good news-bad news effect: Asymmetric processing of objective information about yourself. *Am Econ J Microecon* 3(2):114–138.
35. Reuben E, Sapienza P, Zingales L (2014) How stereotypes impair women's careers in science. *Proc Natl Acad Sci USA* 111(12):4403–4408.
36. Todorov A, Olivola CY, Dotsch R, Mende-Siedlecki P (2015) Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annu Rev Psychol* 66:519–545.
37. Greenwald AG, Nosek BA, Banaji MR (2003) Understanding and using the implicit association test: I. An improved scoring algorithm. *J Pers Soc Psychol* 85(2):197–216.
38. Pinheiro J, Bates D, DebRoy S, Sarkar D, R Development Core Team (2016) *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1–128. Available at CRAN.R-project.org/package=nlme. Accessed June 2, 2016.