nature human behaviour

LETTERS

DOI: 10.1038/s41562-017-0218-y

# Statistically inaccurate and morally unfair judgements via base rate intrusion

Jack Cao[1*], Max Kleiman-Weiner[2] and Mahzarin R. Banaji[1]

**From a statistical standpoint, judgements about an individual are more accurate if base rates about the individual's social group are taken into account[1–4]. But from a moral standpoint, using these base rates is considered unfair and can even be illegal[5–9]. 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 judgements to be both accurate and fair. Despite this alignment, we show that social judgements were inaccurate and unfair. Instead of appropriately setting aside social group differences, participants erroneously relied on them when making judgements 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 non-social judgements. Results also demonstrate how social judgements 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[10–12], these base rates can undermine the rationality and fairness of human judgements.**

Many studies illustrate the importance of base rates for making accurate judgements. 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[1]. 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[2]. 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 judgements[3,4].

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 judgement about an individual will be accurate, as evidenced by decades of research conceptualizing stereotypes as base rates[13–17]. But unlike other base rates, using stereotypes raises serious questions about fairness[5]. Many theories of morality eschew the application of group characteristics to specific individuals because doing so violates individual rights and basic tenets of justice[6,7]. 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[8], nor can they be used to determine insurance premiums in the European Union[9]. 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 judgements 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[18]. 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 judgements about the bottom node. This is an example of the Markov assumption[19], which specifies when variables become conditionally independent of one another (see Supplementary Information for probability calculus). Although previous work has shown that people violate the Markov assumption[20,21], the current work systematically tests base rates from social versus non-social 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 non-social 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 versus non-social base rates fundamentally differ in their influence on human judgements is unclear. On the one hand, both types of base

[1]Department of Psychology, Harvard University, Cambridge, MA 02138, USA. [2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. *e-mail: jackcao@fas.harvard.edu



© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

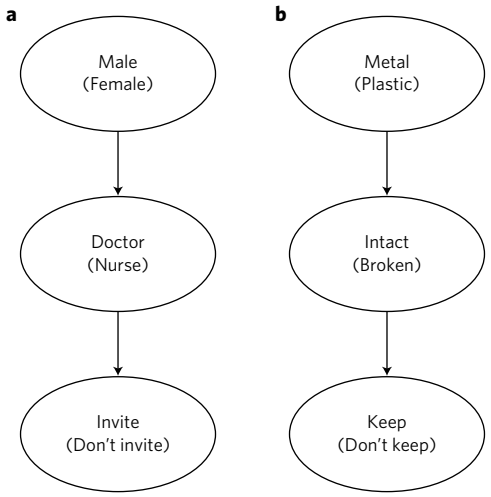


**Fig. 1 | Two Bayesian networks that are identical in structure but differ in whether the base rates are social or non-social. a**, Social base rates. Gender influences an individual's profession, which influences whether he or she will be invited by a charity. **b**, Non-social base rates. A spoon's material influences whether it remains intact or breaks, which influences whether it will be kept by a factory.

rates help simplify a complex world[22,23], which raises possibility that social and non-social knowledge share the same cognitive underpinnings[15,24]. On the other hand, social knowledge is often infused with greater emotion[25] and is differentially structured compared with knowledge about physical objects[26]. Indeed, social neuroscientists have found that thinking about social entities activates a distinct set of brain regions relative to thinking about non-social entities, suggesting that different mechanisms may support these two processes[27,28].

To investigate the potentially unique standing of social base rates vis-à-vis human judgements, 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 versus 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 Fig. 1b. The wording of the scenarios was adapted from ref. [29], which established wording that conveyed a chain network structure. After reading the scenario, each participant made two judgements about the bottom node given knowledge of the top and middle nodes. The only difference between these two judgements was what state of the top node was known. In the spoon scenario, participants judged the likelihood of keeping a broken metal spoon versus a broken plastic spoon, or the likelihood of keeping an intact metal spoon versus an intact plastic spoon. In the gender scenario, participants judged the likelihood of inviting a male nurse versus a female nurse, or the likelihood of inviting a male doctor versus a female doctor. If base rates are properly set aside, both judgements on the 1–7 Likert-type scale (1 = extremely unlikely … 7 = extremely likely) should be the same.

In experiment 1, the Markov assumption was violated, a result that replicates previous work[20,21]. Judgements 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 non-social or social [$F(1, 395) = 63.06$, $P < 0.0001$]. In the spoon scenario, which contained non-social base rates, the plastic spoon was judged less likely to be kept regardless of whether the two spoons in question were broken [$M_{\text{broken metal}} = 4.91$ versus $M_{\text{broken plastic}} = 3.73$; $b = 1.19$, $t(395) = 5.51$, $P < 0.0001$] or intact [$M_{\text{intact metal}} = 6.52$ versus $M_{\text{intact plastic}} = 5.10$; $b = 1.42$, $t(395) = 6.53$, $P < 0.0001$]. But in the gender scenario, which contained social base rates, judgements erroneously relied upon group differences. A male nurse was judged less likely to be invited than a female nurse [$M_{\text{male nurse}} = 3.70$ versus $M_{\text{female nurse}} = 5.58$; $b = -1.88$, $t(395) = -8.69$, $P < 0.0001$], but a female doctor was judged less likely to be invited than a male doctor [$M_{\text{male doctor}} = 6.18$ versus $M_{\text{female doctor}} = 4.36$; $b = 1.81$, $t(395) = 8.16$, $P < 0.0001$]. Despite having the opportunity to be both statistically accurate and morally fair, social judgements broke with Bayesian rationality and with tenets of fairness. The small main effect of profession (nurse versus doctor) in the social condition raises the possibility that participants may not have comprehended the scenarios. We ruled out this possibility by replicating experiment 1 and including a comprehension check (Supplementary Figs. 1 and 2).

Experiment 2 sought to replicate the findings and ensure that they were not specific to non-social base rates about spoons or to social base rates about gender. These base rates were tested once more alongside other non-social 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 versus white), nationality (American versus foreign), and age (young versus 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 non-social domains. Alarms emanate from inanimate physical objects, but burglary is an activity perpetrated by one

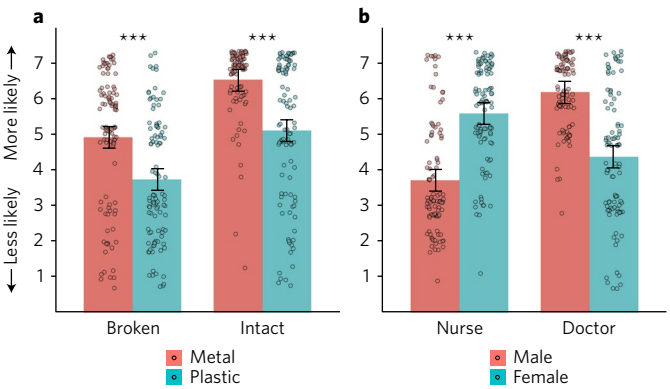


**Fig. 2 | The erroneous influence of base rates in experiment 1 ($N = 399$). a**,**b**, Participants' average judgements when non-social (**a**) versus social (**b**) base rates should not have been used. Each participant was randomly assigned to make two judgements on a 1 to 7 scale (for example, broken metal spoon versus broken plastic spoon). All possible pairs of judgements a participant could have been randomly assigned to make are on the *x*-axes. Statistical tests compare the means of each pair of judgements. Points show the distribution of judgements. Error bars are 95% confidence interval. ***$P < 0.001$.



© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

group (burglars) against another (tenants and homeowners). This blending raises the possibility that social and non-social 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 non-social or social (Supplementary Fig. 5). When base rates were non-social, judgements 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 judgements that upheld neither Bayesian rationality nor tenets of fairness—even though upholding both was possible.

When two individuals differed in gender but not in profession, participants' judgements erroneously relied upon gender differences. As a robustness test of this effect, participants in experiment 3 made judgements about two individuals who differed in profession but not in gender. That is, participants judged how likely a male nurse versus a male doctor were to be invited, or they judged how likely a female nurse versus female doctor were to be invited. Without a contrast gender, perhaps participants would base their judgements 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 non-social 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 [$M_{\text{male nurse}} = 3.55$ versus $M_{\text{female nurse}} = 5.81$; $b = -2.25$, $t(394) = -11.49$, $P < 0.0001$], but a female doctor was judged less likely to be invited than a male doctor [$M_{\text{male doctor}} = 6.25$ versus $M_{\text{female doctor}} = 3.82$; $b = 2.43$, $t(394) = 12.40$, $P < 0.0001$]. Again, social judgements upheld neither Bayesian rationality nor tenets of fairness, even though both accuracy and fairness were achievable.

Having established the generalizability and robustness of the effect, we next set out to correct it. Can social judgements 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 judgements to be both accurate and fair, three potential problems with experiments 1–3 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[30]—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 versus females. Likewise, the spoon scenario explicitly referenced different breakage rates between metal versus plastic spoons. These explicit base rate references may have led participants to assume that base rates were relevant to judgements. To prevent participants from following this Gricean maxim of relvance[31]—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, judgements should have been about the bottom node given knowledge of the top and middle nodes. However, participants may have done the opposite by making judgements about the top and middle nodes given knowledge of the bottom node. To prevent participants from committing this inverse fallacy[32]—misinterpreting the judgement 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 judgement needed to be made.

When all three of these strategies were implemented simultaneously, social judgements were both accurate and fair (Fig. 3). Although non-social judgements were still erroneously influenced by a spoon's material, a person's gender no longer substantially influenced participants' social judgements [$F(1, 407) = 60.15$, $P < 0.0001$]. With respect to who would be invited, parity was achieved between a male nurse and female nurse [$M_{\text{male nurse}} = 4.46$ versus $M_{\text{female nurse}} = 4.61$; $b = -0.15$, $t(407) = -1.32$, $P = 0.19$] and between a female doctor and male doctor [$M_{\text{male doctor}} = 5.76$ versus $M_{\text{female doctor}} = 5.51$; $b = 0.25$, $t(407) = 2.29$, $P = 0.02$]. Gender–profession base rates were kept at bay, resulting in social judgements that were consistent with Bayesian rationality and with tenets of fairness.

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 judgements improved somewhat but still erroneously relied upon group differences that should have been disregarded, showing that this bias 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 judgements (Supplementary Fig. 10). Although future research is needed to confirm this finding, it is consistent with Fiske and Neuberg's model of impression formation[10]: 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 judgements.

The results of experiment 4 also underscore the extensive work required to construct a representation of the task that enables social judgements to be accurate and fair. Although the tasks in experiments 1–3 lack the remedies that together eliminated the bias, these earlier tasks faithfully represent the conditions under which social

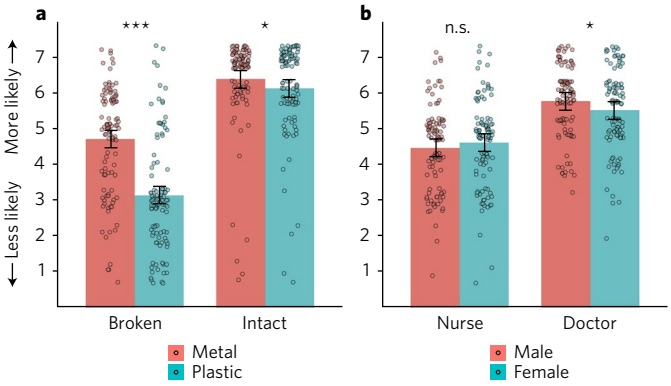


**Fig. 3 | Social base rates no longer influenced judgments in experiment 4 (*N* = 411).** Participants' average judgements when non-social (**a**) versus social (**b**) base rates should not have been used. Each participant was randomly assigned to make two judgements on a 1 to 7 scale (for example, broken metal spoon versus broken plastic spoon). All possible pairs of judgements a participant could have been randomly assigned to make are on the *x*-axes. Statistical tests compare the means of each pair of judgements. Points show the distribution of judgements. Error bars are 95% confidence interval. ***$P < 0.001$, *$P < 0.05$, n.s. $P > 0.05$. When all three strategies were implemented, the effect remains in non-social judgements, but social judgements become accurate and fair.



© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

judgements are typically made. In everyday life, disparate outcomes between social groups are common[11]; base rates about these groups are exaggerated in stereotype knowledge[12]; and stereotyped people are not always uniquely individuated[10]. However, the presence of any one of these features can lead to the mistaken use of gender, race, nationality, or age in social judgements.

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 judgements of statistical likelihood[1–4]. 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 judgements fell short on both accounts.

This bias was observed not only in participants' average judgements, but also in contour plots that depict which 1–7 likelihood judgements 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 judgements. These contour plots also refute an alternative account, namely that stereotype-incongruent targets (for example, male nurses, female doctors) may have decreased confidence or induced confusion among participants[33]. If this were the case, then there should be a high density of participants who gave judgements at the low end (1) or midpoint (4) of the scale. However, 1s or 4s were hardly observed for judgements about stereotype-incongruent targets. Instead, the distribution of judgements is consistent with the erroneous use of base rates.

These contour plots also reveal a different error pattern for non-social judgements, relative to social judgements. 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 judgements. However, many participants were able avoid this error, as there are clusters of participants whose judgements lie on the 45-degree identity line (Supplementary Fig. 13). These findings further suggest important differences in how social versus non-social base rates influence human judgements 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 non-social base rates. Consequently, the likely medical professions of men versus women, for example, may hold considerable sway over social judgements, leading to robust judgements that favour female nurses over male nurses but male doctors over female doctors. However, strategies that virtually eliminated base rate intrusion in social judgements did not have the same ameliorating effect on non-social judgements (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 non-social base rates. Knowledge of social group differences raises questions of fairness, perhaps motivating people to disregard this knowledge if the judgement task is constructed as it was in experiment 4. Current theories of causal reasoning are agnostic to semantic content[20,21], 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 versus non-social 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 judgement also becomes a fair judgement. 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 judgements about other people.

## Methods

**Participants, sample size and informed consent.** All participants were recruited from Amazon Mechanical Turk. Sample sizes of approximately 100 participants per cell were determined a priori based on previous research (see Supplementary Information for demographic information and sample size, mean and s.d. 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 non-social scenario (see Supplementary Information for stimuli). After reading the scenario, each participant made two judgements about the bottom node given knowledge of the top and middle nodes (Fig. 1). The only difference between these two judgements was what state of the top node was known. In the non-social scenario, about spoons, participants were randomly assigned to judge the likelihood of keeping a broken metal spoon versus a broken plastic spoon, or the likelihood of keeping an intact metal spoon versus an intact plastic spoon. In the social scenario, about gender, participants were randomly assigned to judge the likelihood of inviting a male nurse versus a female nurse, or the likelihood of inviting a male doctor versus a female doctor. All judgements 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 judgements. Social and non-social content were varied in experiment 2. In experiment 3, participants again made two judgements, 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 the nlme package in R[34]. The three-way interaction between base rate (non-social versus social), the middle node in the chain network (broken/nurse versus intact/doctor), and the top node in the chain network (plastic/female versus 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[35]. 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.

**Code availability.** All code is available on the Open Science Framework (https://osf.io/htpzr).

**Data availability.** All data are available on the Open Science Framework (https://osf.io/htpzr).

Received: 22 December 2016; Accepted: 17 August 2017;
Published online: 02 October 2017

## References

1. Eddy, D. M. in *Judgment Under Uncertainty: Heuristics and Biases* (eds Kahneman, D., Slovic, P. & Tversky, A.) 249–267 (Cambridge Univ. Press, Cambridge, 1982).
2. Kahneman, D. & Tversky, A. On the psychology of prediction. *Psychol. Rev.* **80**, 237–251 (1973).
3. Bar-Hillel, M. The base-rate fallacy in probability judgments. *Acta Psychol.* **44**, 211–233 (1980).
4. Tversky, A. & Kahneman, D. Judgment under uncertainty: heuristics and biases. *Science* **185**, 1124–1131 (1974).
5. Cao, J. & Banaji, M. R. The base rate principle and the fairness principle in social judgment. *Proc. Natl Acad. Sci. USA* **113**, 7475–7580 (2016).
6. Rawls, J. *Justice as Fairness: A Restatement* (Harvard Univ. Press, Cambridge, 2001).
7. Dworkin, R. *Sovereign Virtue: The Theory and Practice of Equality* (Harvard Univ. Press, Cambridge, 2000).



© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

8. Koehler, J. in *Handbook of Psychology and Law* (eds Kagehiro, D. & Laufer, W.) 167–184 (Springer, New York, 1992).
9. *Test-Achats v. Council of Ministers* (European Court of Justice, 2011).
10. Fiske, S. & Neuberg, S. 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 (1990).
11. Moss-Racusin, C., Dovidio, J., Brescoll, V., Graham, M. & Handelsman, J. Science faculty's subtle gender biases favor male students. *Proc. Natl Acad. Sci. USA* **109**, 16474–16479 (2012).
12. Cheryan, S., Plaut, C., Davies, P. & Steele, C. Ambient belonging: how stereotypical cues impact gender participation in computer science. *J. Pers. Soc. Psychol.* **97**, 1056–1060 (2009).
13. Locksley, A., Borgida, E., Brekke, N. & Hepburn, C. Sex stereotypes and social judgment. *J. Pers. Soc. Psychol.* **39**, 821–831 (1980).
14. Rasinksi, K. A., Crocker, J. & Hastie, R. Another look at sex stereotypes and social judgments: an analysis of the perceiver's use of subjective probabilities. *J. Pers. Soc. Psychol.* **49**, 317–326 (1985).
15. Hamilton, D. L. *Cognitive Processes in Stereotyping and Intergroup Behavior* (Erlbaum, Hillsdale, 1981).
16. Krosnick, J. A., Li, F. & Lehman, D. R. Conversational conventions, order of information acquisition, and the effect of base rates on individuating social information on social judgments. *J. Pers. Soc. Psychol.* **59**, 1140–1152 (1990).
17. Jussim, L. *Social Perception and Social Reality: Why Accuracy Dominates Bias and Self-Fulfilling Prophecy* (Oxford Univ. Press, Oxford, 2012).
18. Pearl, J. *Causality: Models, Reasoning, and Inference* (Cambridge Univ. Press, Cambridge, 2000).
19. Hausman, D. M. & Woodard, J. Independence, invariance, and the causal Markov condition. *Brit. J. Phil. Sci.* **50**, 521–583 (1999).
20. Rottman, B. M. & Hastie, R. Reasoning about causal relationships. Inferences on causal networks. *Psychol. Bull.* **140**, 109–139 (2014).
21. Rehder, B. Independence and dependence in human causal reasoning. *Cogn. Psychol.* **72**, 54–107 (2014).
22. Murphy, G. *The Big Book of Concepts* (MIT, Cambridge, 2002).
23. Medin, D. & Smith, E. Concepts and concept formation. *Annu. Rev. Psychol.* **35**, 113–138 (1984).
24. Banaji, M. & Bhaskar, R. in *Memory, Brain, and Belief* (eds Schacter, D. & Scarry, E.) 139–175 (Harvard Univ. Press, Cambridge, 1999).
25. Norris, C., Chen, E., Zhu, D., Small, S. & Cacioppo, J. The interaction of social and emotional processes in the brain. *J. Cogn. Neurosci.* **16**, 1818–1829 (2004).
26. Wattenmaker, W. Knowledge structures and linear separability: integrating information in object and social categorization. *Cogn. Psychol.* **28**, 273–328 (1995).
27. Contreras, J., Banaji, M. & Mitchell, J. Dissociable neural correlations of stereotypes and other forms of semantic knowledge. *Socl. Cogn. Affect. Neurosci.* **7**, 764–770 (2012).
28. Mitchell, J., Heatherton, T. & Macrae, C. Distinct neural systems subserve person and object knowledge. *Proc. Natl Acad. Sci. USA* **99**, 15238–15243 (2002).
29. Krynski, T. & Tenenbaum, J. The role of causality in judgment under uncertainty. *J. Exp. Psychol. Gen.* **126**, 430–450 (2007).
30. Kay, A. et al. Inequality, discrimination, and the power of the status quo: direct evidence for a motivation to view what is as what should be. *J. Pers. Soc. Psychol.* **97**, 421–434 (2009).
31. Grice, H. P. in *Syntax and Semantics* (eds Cole, P. and Morgan, J.) 41–58 (Academic Press, New York, 1975).
32. Villejoubert, G. & Mandel, D. The inverse fallacy: an account of deviations from Bayes's theorem and the additivity principle. *Mem. Cogn.* **30**, 171–178 (2002).
33. Rottman, B. M. & Hastie, R. Do people reason rationally about causally related events? Markov violations, weak inferences, and failures in explaining away. *Cogn. Psychol.* **87**, 88–134 (2016).
34. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team nlme: linear and nonlinear mixed effects models. R package v. 3.1-131 (2017).
35. Bates, D., Maechler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).


## Acknowledgements

This work was supported by NSF Graduate Research Fellowships to J.C. and M.K.W.; an Inequality and Social Policy fellowship from Harvard University Kennedy School of Government to J.C.; a Hertz Fellowship to M.K.W. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We are grateful to B. Rehder and S. Gershman for helpful comments and to K. Morehouse for research assistance.

## Author contributions

J.C., M.K.W. and M.R.B. designed research. J.C. performed research and analysed data. J.C., M.K.W. and M.R.B. wrote the paper.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41562-017-0218-y.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Correspondence and requests for materials** should be addressed to J.C.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

nature research

Corresponding author(s): Jack Cao (jackcao@fas.harvard.edu)

☐ Initial submission ☐ Revised version ☒ Final submission

# Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

## Experimental design

1. Sample size

   Describe how sample size was determined.

   Large sample sizes of approximately 100 participants per cell were specified a priori based on past research.

2. Data exclusions

   Describe any data exclusions.

   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. This exclusion criterion was not pre-specified.

   Data from Experiments 4, 6, and 7 were collected simultaneously. Five participants were excluded for not completing the procedure. This exclusion criterion was pre-specified.

3. Replication

   Describe whether the experimental findings were reliably reproduced.

   All replication attempts were successful. They are included in the paper as separate experiments.

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   Participants were randomly assigned to condition using Qualtrics.

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   All data were collected on Amazon Mechanical Turk and no analyses for an experiment began until after data collection was complete. During data collection, we did not know which conditions participants were assigned to. During analysis, we knew which conditions participants were assigned to.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

| n/a | Confirmed |
|---|---|
| ☐ | ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.) |
| ☐ | ☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ A statement indicating how many times each experiment was replicated |
| ☐ | ☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section) |
| ☐ | ☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons |
| ☐ | ☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted |
| ☐ | ☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range) |
| ☐ | ☒ Clearly defined error bars |

*See the web collection on statistics for biologists for further resources and guidance.*

## Software

Policy information about availability of computer code

7. Software

| | |
|---|---|
| Describe the software used to analyze the data in this study. | All analyses were conducted using R statistical computing, specifically the the nlme package by Pinheiro et al. (2017) and the lme4 package by Bates et al. (2015). |

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

## Materials and reagents

Policy information about availability of materials

8. Materials availability

| | |
|---|---|
| Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company. | No unique materials were used. All materials used are available on the Open Science Framework (see manuscript for URL). |

9. Antibodies

| | |
|---|---|
| Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species). | No antibodies were used. |

10. Eukaryotic cell lines

| | |
|---|---|
| a. State the source of each eukaryotic cell line used. | No eukaryotic cell lines were used. |
| b. Describe the method of cell line authentication used. | No eukaryotic cell lines were used. |
| c. Report whether the cell lines were tested for mycoplasma contamination. | No eukaryotic cell lines were used. |
| d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use. | No commonly misidentified cell lines were used. |

## Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

| | |
|---|---|
| Provide details on animals and/or animal-derived materials used in the study. | No animals were used. |

Policy information about studies involving human research participants

## 12. Description of human research participants

| | |
|---|---|
| Describe the covariate-relevant population characteristics of the human research participants. | All participants were recruited from Amazon Mechanical Turk. Approximately 50% of all participants were female and the mean age was in the early to mid 30s, depending on experiment. The standard deviation for age was approximately 11 years, again depending on experiment. We provide exact gender distributions and age information in the Supplementary Information. |