Supplemental Materials for:

People make the Bayesian judgment they criticize in others

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

[1]Harvard University, Department of Psychology

[2]Massachusetts Institute of Technology, Department of Brain & Cognitive Sciences

[*]Corresponding author

jackcao@fas.harvard.edu

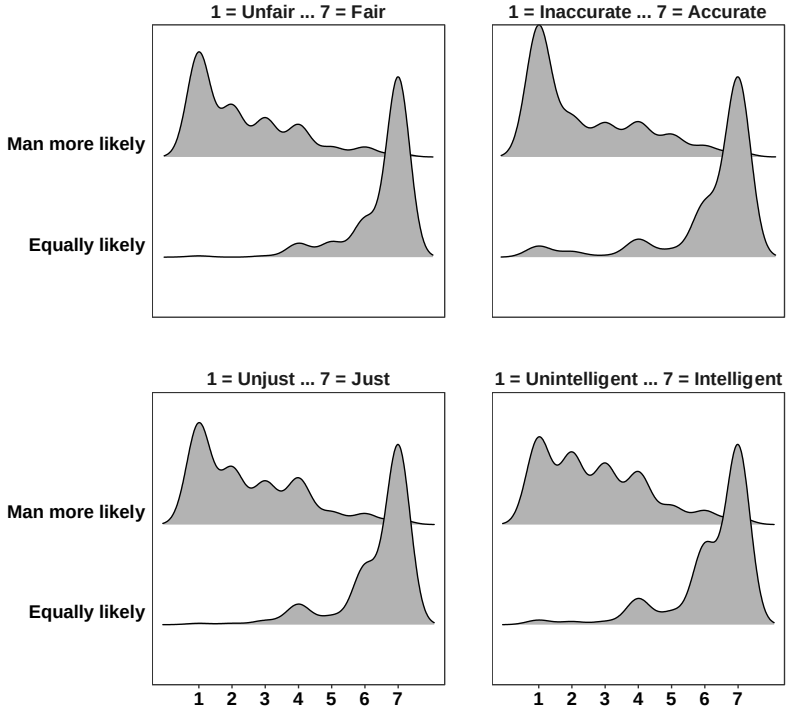
Table S1. Study 2. Item means and standard errors.

	Butcher Mean (SE)	*Firefighter* Mean (SE)	*Construction Worker* Mean (SE)
Fair	3.28 (0.13)	3.19 (0.13)	3.62 (0.13)
Just	3.35 (0.12)	3.21 (0.12)	3.74 (0.13)
Accurate	3.70 (0.14)	4.09 (0.15)	4.09 (0.13)
Intelligent	3.59 (0.12)	3.67 (0.12)	3.85 (0.12)

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).

	Man more likely	**Equally likely**
Fair	2.34 (0.11)	6.51 (0.07)
Just	2.51 (0.11)	6.39 (0.08)
Accurate	2.38 (0.12)	6.21 (0.11)
Intelligent	2.73 (0.11)	6.23 (0.09)

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



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_{age} = 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 *SE*s) 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 $\alpha = 0.93$, $M_{composite} = 2.49$, $SE = 0.10$, 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 $\alpha = 0.85$, $M_{composite} = 6.36$, $SD = 0.07$, 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).

	Man more likely	Equally likely
Fair	2.38 (0.11)	6.51 (0.07)
Just	2.39 (0.11)	6.47 (0.07)
Accurate	2.42 (0.12)	6.28 (0.09)
Intelligent	2.77 (0.11)	6.18 (0.08)

Fig. S2. Additional study: costly punishment. Distributions of evaluations of Person X.

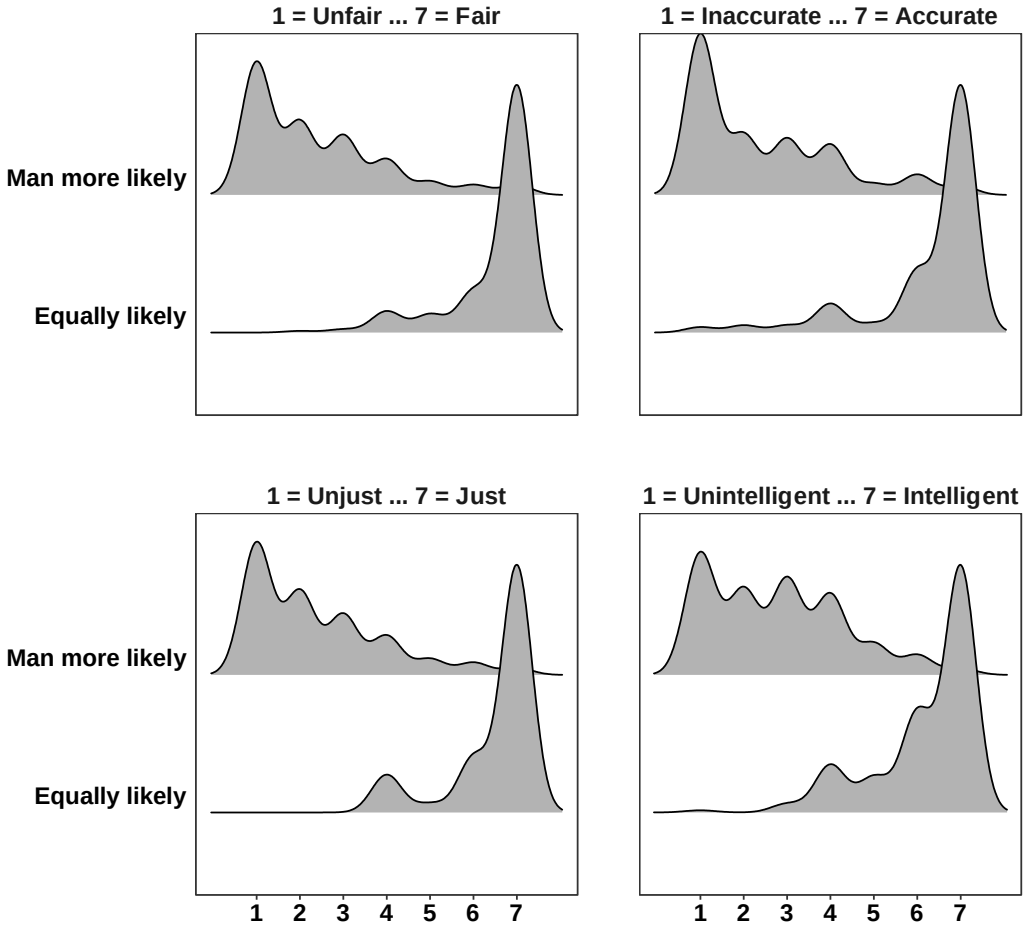
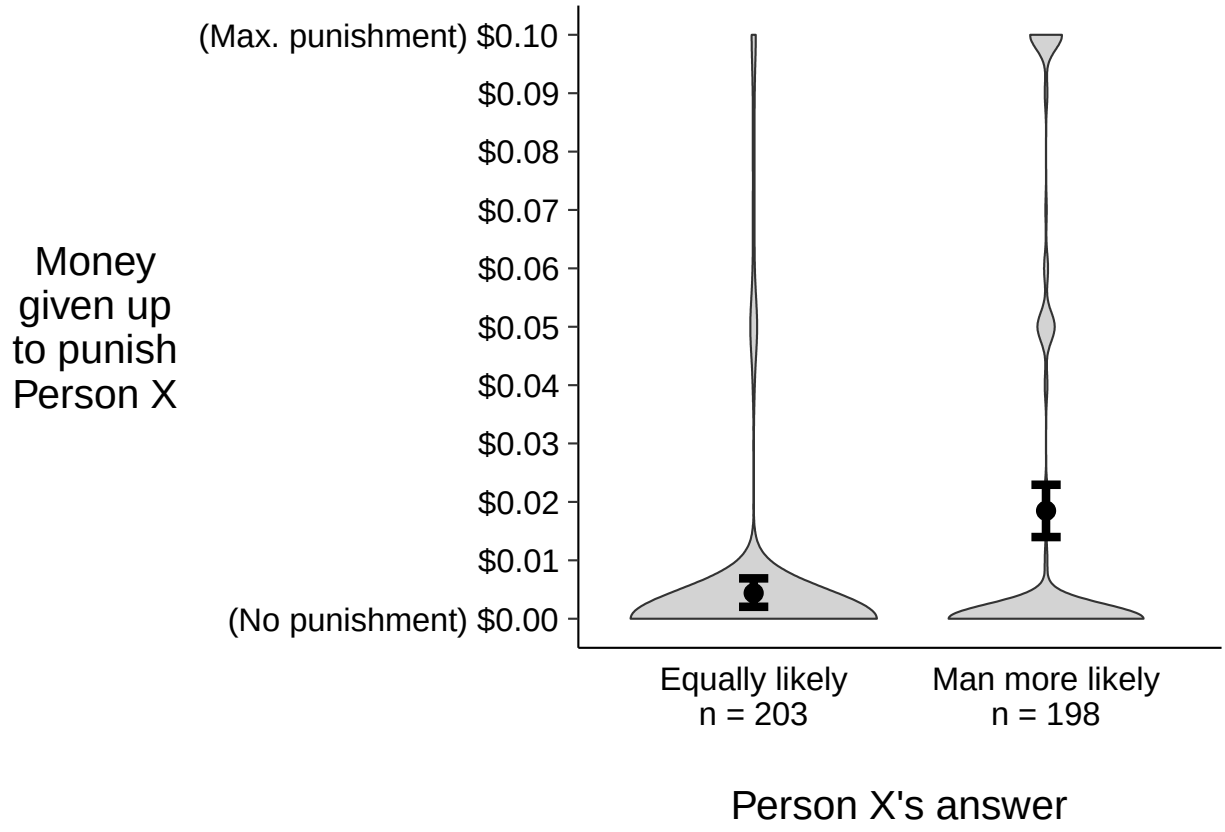


Fig. S3. Additional study: costly punishment. Average amounts given up to punish Person X. Errors bars are 95% *CI*s. Violin plots display the distribution of amounts given up in each condition.



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_{age} = 34.00 years, *SD* = 10.04; 105 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 α = 0.93, $M_{composite}$ = 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_{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.

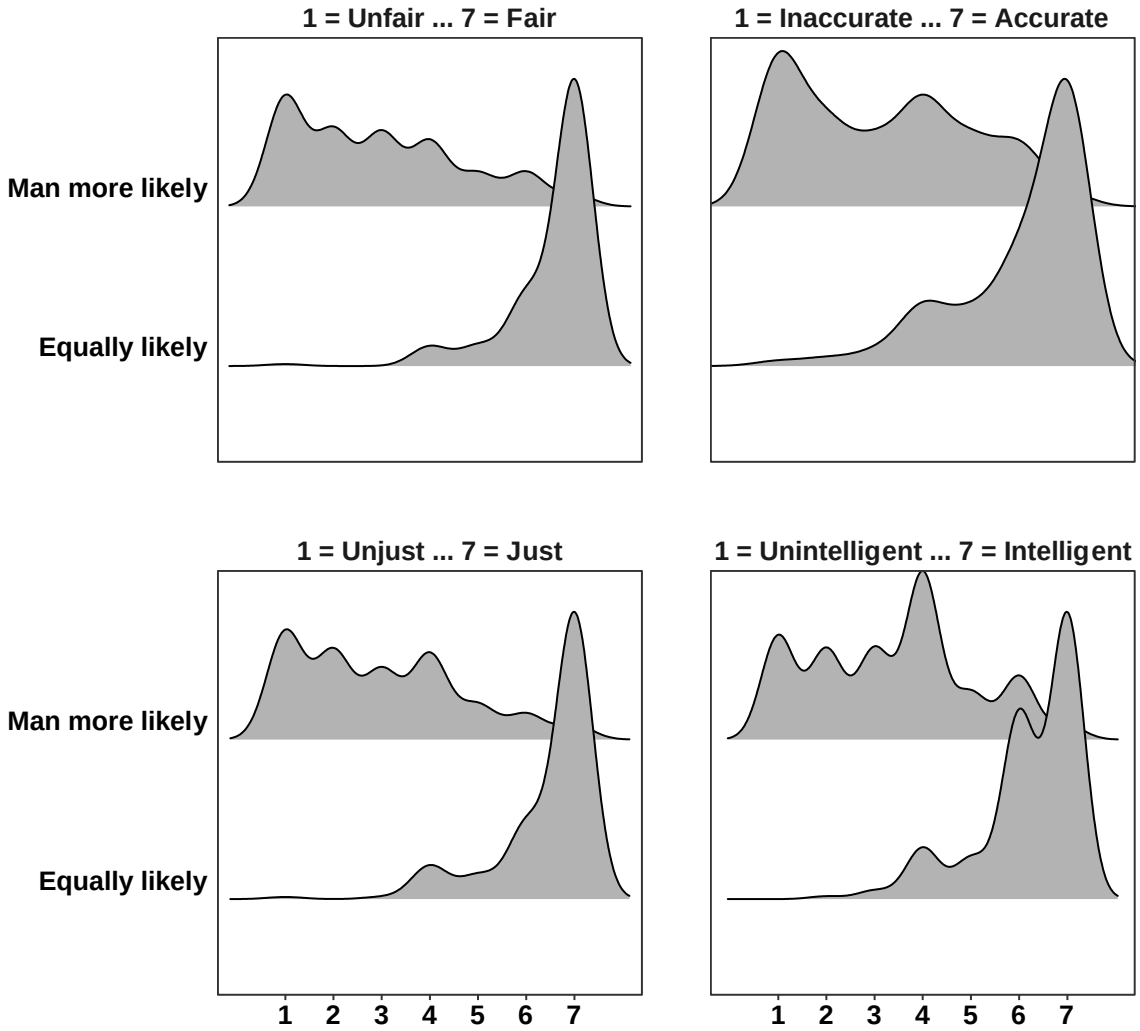
Results. The results were replicated. A majority of participants, 81%, agreed with the egalitarian judgment that the two percentages are equivalent, whereas 15% agreed with the Bayesian judgment that the two percentages differ in that the man is more likely to be the pilot, and 4% agreed that the two 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 *SE*s) 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 $\alpha = 0.93$, $M_{composite} = 3.11$, $SD = 0.11$, 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 $\alpha = 0.85$, $M_{composite} = 6.30$, $SD = 0.06$, 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, $n = 201$).

	Man more likely	**Equally likely**
Fair	2.96 (0.12)	6.55 (0.06)
Just	2.96 (0.12)	6.43 (0.07)
Accurate	3.15 (0.13)	6.05 (0.10)
Intelligent	3.37 (0.12)	6.18 (0.07)

166 **Fig. S4.** Additional study showing conceptual replication negative evaluations of Person X.
167 Distributions of evaluations of Person X.



168
169
170
171
172
173
174
175
176
177
178

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, $Median_{Man}$ = -2.36 vs. $Median_{Woman}$ = -2.77, Wilcoxon P = 0.05, r = 0.11. Moreover, the log of these likelihoods 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, $M_{Model\ Posterior,\ Man}$ = 24.3% vs. $M_{Model\ Posterior,\ Woman}$ = 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, $M_{Reported\ Posterior,\ Man}$ = 31.6% vs. $M_{Reported\ Posterior,\ Woman}$ = 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, $M_{Model\ Posterior,\ Man}$ = 24.3% vs. $M_{Reported\ Posterior,\ Man}$ = 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, $M_{Model\ Posterior,\ Woman}$ = 7.4% vs. $M_{Reported\ Posterior,\ Woman}$ = 13.1%, b = -0.06, $t(1780)$ = -3.02, P = 0.003, r = 0.07. Error bars are 95% *CI*s.

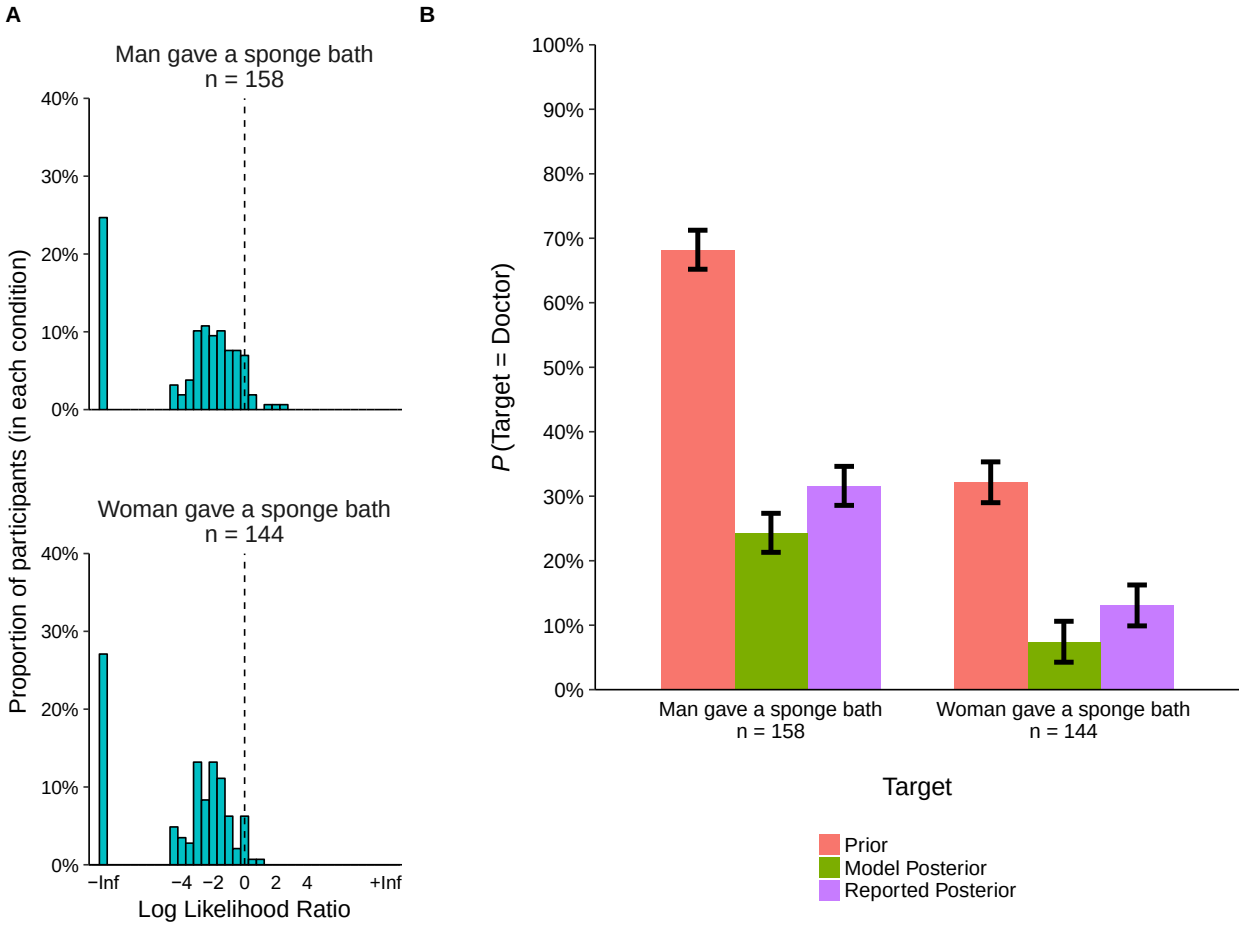


Fig. S6. Study 4: CPR conditions. **A.** Minimal differences in likelihood ratios were observed between participants who learned that the man vs. woman had performed CPR, $Median_{Man}$ = -0.19 vs. $Median_{Woman}$ = -0.18, Wilcoxon $P = 0.66$, $r = 0.03$. Moreover, the log of these likelihood ratios 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, $ts(1780) < |3.44|$, $Ps > 0.0006$, $rs < 0.09$. Error bars are 95% *CI*s.

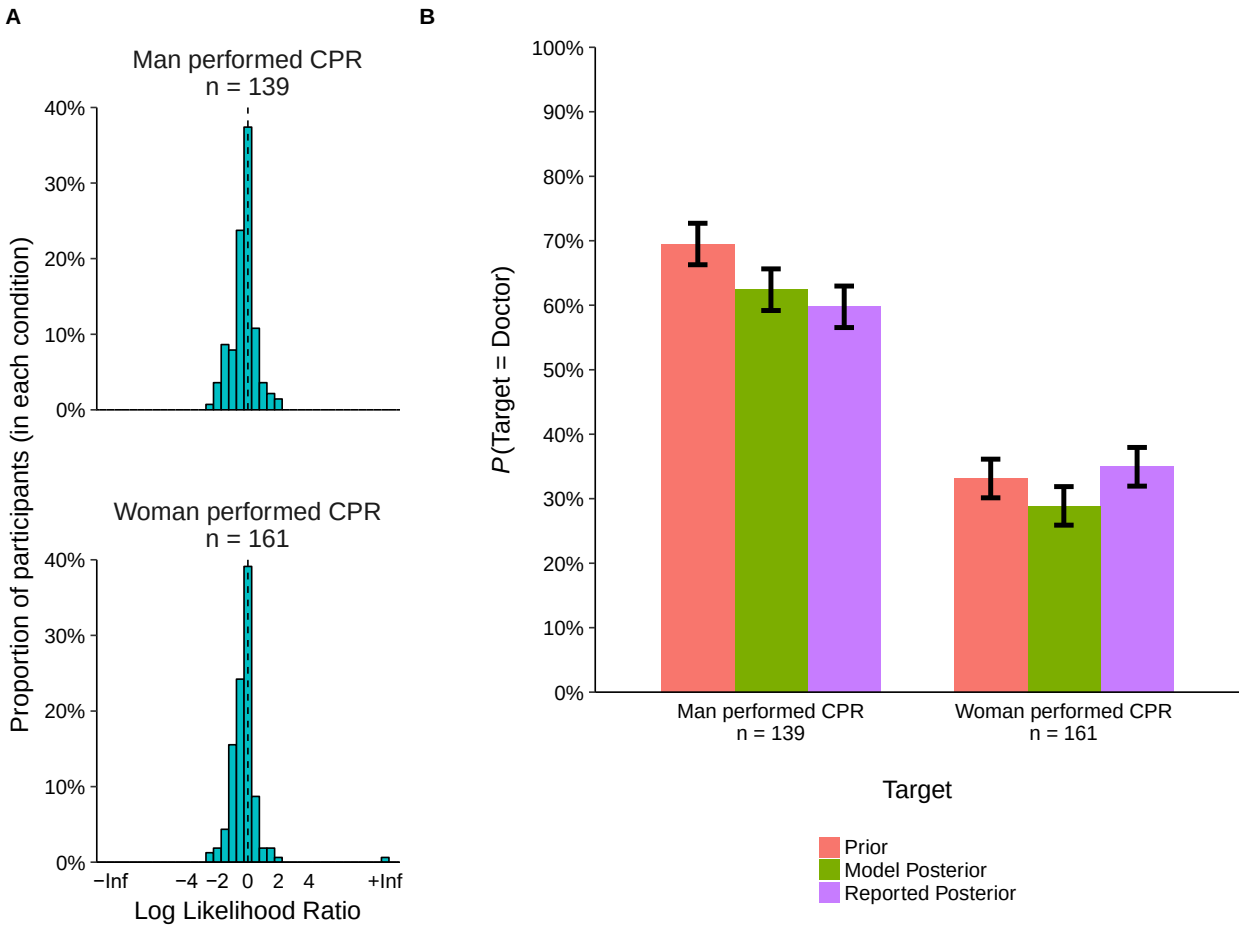


Fig. S7. Study 4: surgery conditions. **A.** The correspondence between model and reported posteriors is present at the level of the individual participant. By subtracting each participant's model posterior from his or her reported posterior, we calculate an accuracy score for each participant, with 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, $B = 0.25$, $t(189) = 3.67$, $P = 0.0003$, $r = 0.26$.

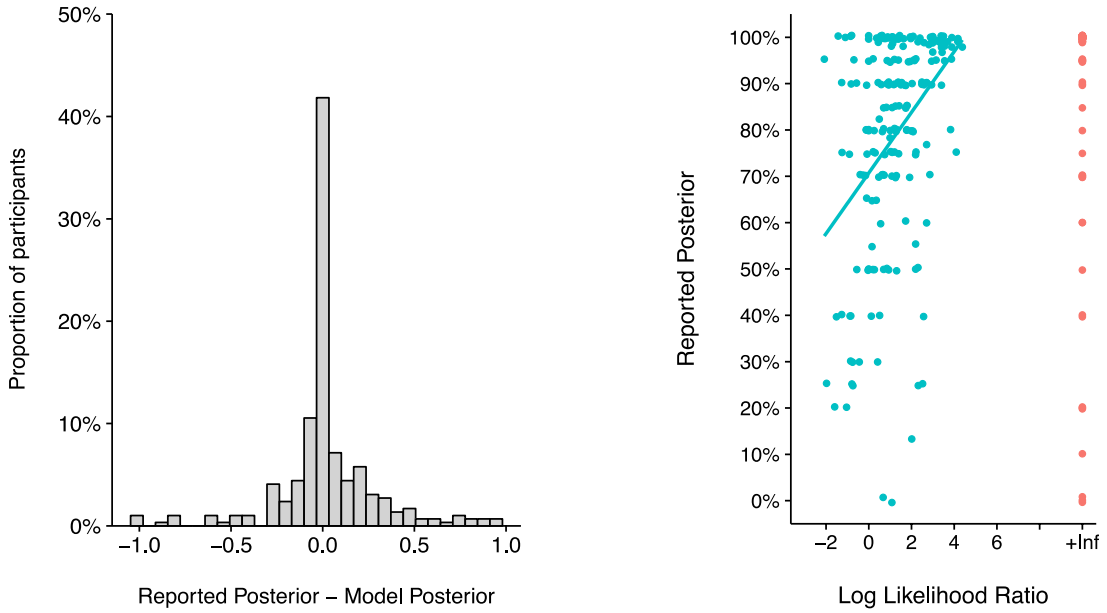
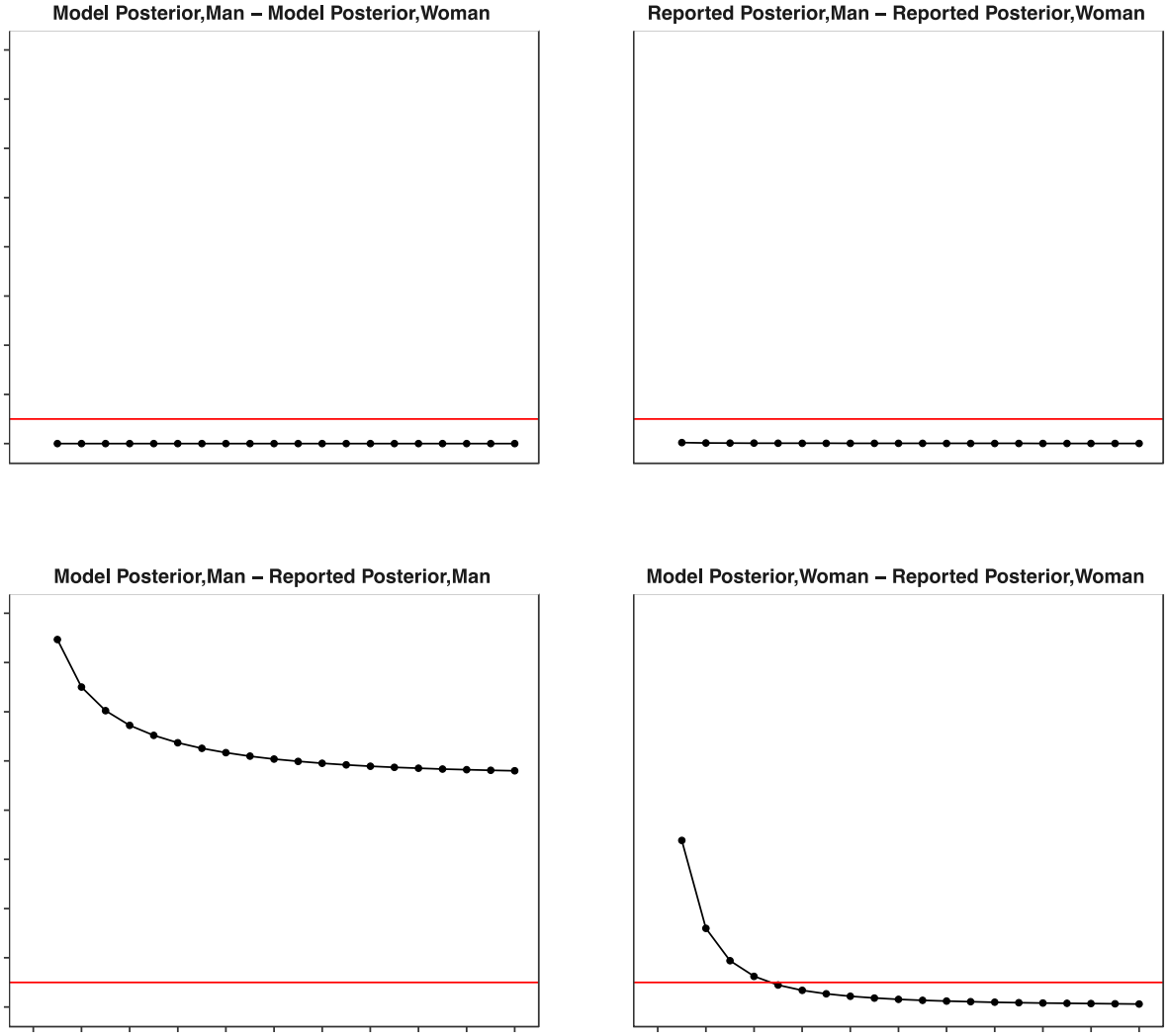


Fig. S8. Study 4: surgery conditions. The statistical significance of the four critical comparisons is robust to the choice of adjustment factor when participants' probability judgments are logit transformed. The adjustment factor is necessary to avoid logit transforming probabilities of 0 or 1. Each 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.
v. The man went through a special line at airport security.

263 vi. The woman went through a special line at airport security.

264

265 Communicating with air traffic control was chosen because it is highly diagnostic of the
266 person being a pilot. Serving beverages was chosen because it is highly diagnostic of the person
267 being a flight attendant. Going through a special line at airport security was chosen because it is
268 relatively non-diagnostic of profession, as both pilots and flight attendants do this. For the
269 primary analysis, only the first two conditions (i and ii, communicated with air traffic control)
270 are discussed. Data from the other four conditions (iii – vi) are presented in Figs. S9-S10.
271 Participants also estimated two likelihoods, the likelihood of observing the datum given
272 the hypothesis that the target they learned about is the pilot as well as the likelihood of
273 observing the datum given the hypothesis that the target they learned about is the flight
274 attendant. If a participant learned that the woman had communicated with air traffic control
275 during a flight, that participant estimated the percentage of female pilots who communicate
276 with traffic control during a flight, as well as the percentage of female flight attendants who
277 communicate with traffic control during a flight. If a participant learned that the man had
278 communicated with air traffic control during a flight, that participant answered the same two
279 questions except about male pilots and male flight attendants. As before in Study 4 of the main
280 text, each participant's priors and likelihoods were entered into Bayes' rule to compute a model
281 posterior, which was then compared against the posterior the participant had reported.

282

283

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_{Man} = 71.2% vs. M_{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, $Median_{Man}$ = 2.23 vs. $Median_{Woman}$ = 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, $M_{Model\ Posterior,\ Man}$ = 90.8% vs. $M_{Model\ Posterior,\ Woman}$ = 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, $M_{Reported\ Posterior,\ Man}$ = 85.8% vs. $M_{Reported\ Posterior,\ Woman}$ = 67.3%, b = 0.18, $t(933) = 7.73$,

306 $P < 0.0001$, $r = 0.25$. Further replicating previous results, small differences were observed
307 between model posteriors and reported posteriors among participants who learned that the
308 man had communicated with air traffic control, $M_{\text{Model Posterior, Man}} = 90.8\%$ vs. $M_{\text{Reported Posterior, Man}}$
309 $= 85.8\%$ %, $b = 0.05$, $t(1866) = 2.59$, $P = 0.01$, $r = 0.06$, and among participants who had learned
310 that the woman had communicated with air traffic control, $M_{\text{Model Posterior, Woman}} = 63.0\%$; vs.
311 $M_{\text{Reported Posterior, Woman}} = 67.3\%$, $b = -0.04$, $t(1866) = -2.24$, $P = 0.03$, $r = 0.05$. So once again, the
312 posteriors reported by participants were close to the posteriors they should have reported
313 according to Bayesian rationality (Fig. S11B).

314 Additional analyses show a) this close correspondence at the level of the individual
315 participant, b) the sensitivity of reported posteriors to likelihood ratios, and c) that the critical
316 comparisons are robust when participants' probability judgments are logit transformed with a
317 wide range of adjustment factors (Figs. S12-S13). In sum, a man who communicated with air
318 traffic control during a flight was judged more likely to be a pilot than a woman who exhibited
319 the same behavior.

320

321

322

323

324

325

326

327

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, $Median_{Man}$ = -Inf vs. $Median_{Woman}$ = -Inf, Wilcoxon P = 0.67, r = 0.02. Moreover, the log of these likelihoods 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% *CI*s.

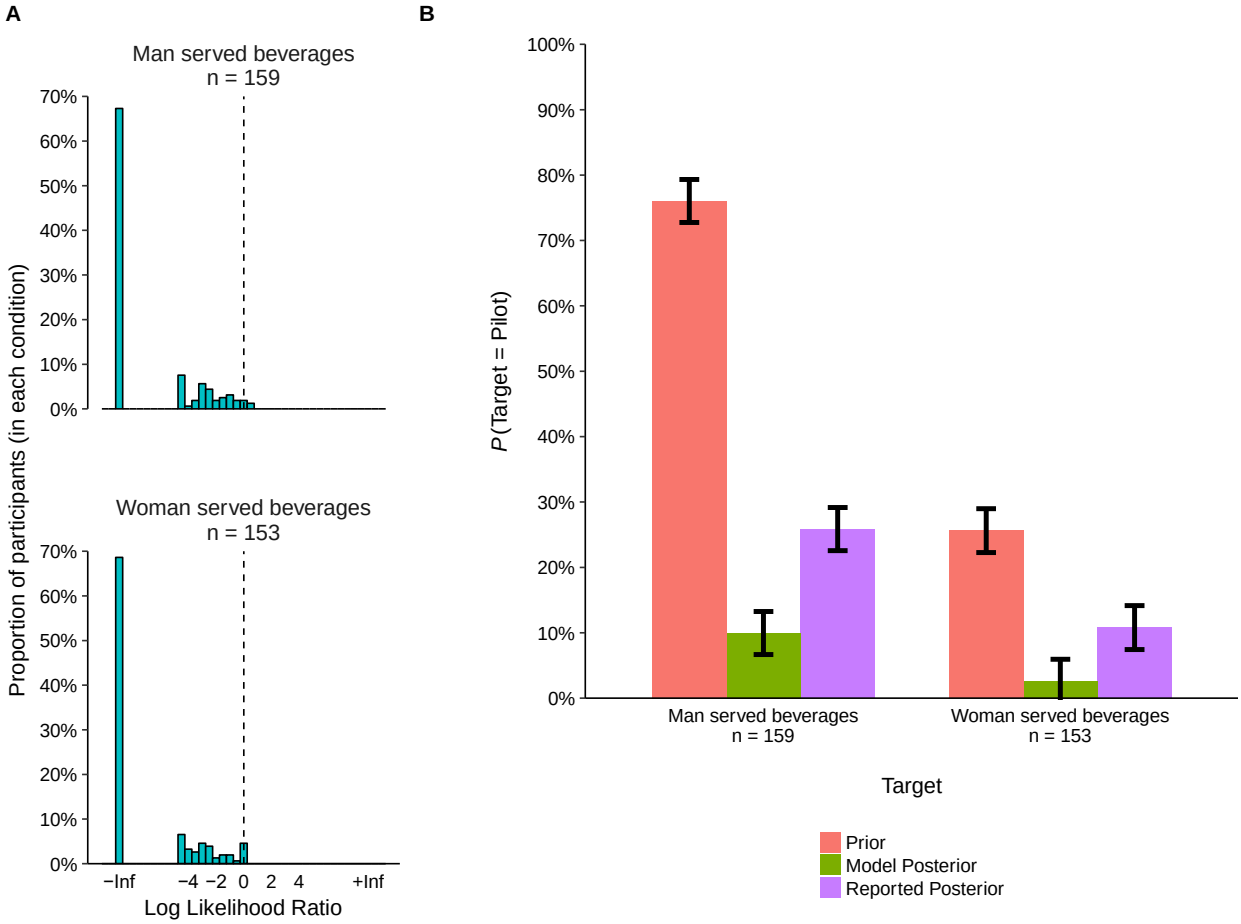


Fig. S10. Additional study that conceptually replicates Bayesian judgments: special line conditions. **A.** Minimal differences in likelihood ratios were observed between participants who learned that the man vs. woman had gone through a special line at airport security, $Median_{Man}$ = 0 vs. $Median_{Woman}$ = 0, Wilcoxon P = 0.01, r = 0.14. Moreover, the log of these likelihoods 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% *CI*s.

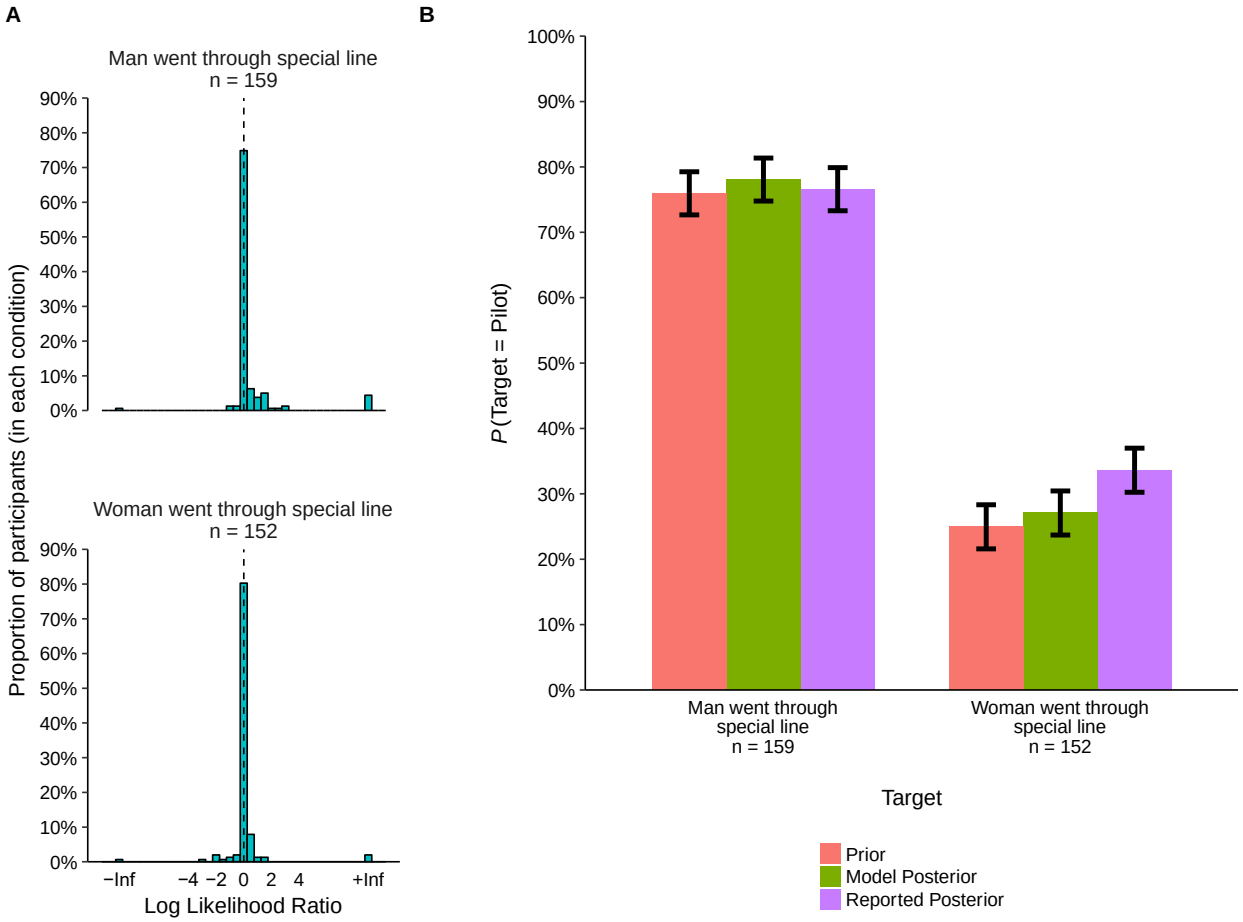


Fig. S11. Additional study that conceptually replicates Bayesian judgments: communicated with air traffic control (ATC) conditions. **A.** Distribution of likelihood ratios (log scaled) in each condition. **B.** Average judgments among participants in each condition. Priors indicate judgments before 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% *CI*s.

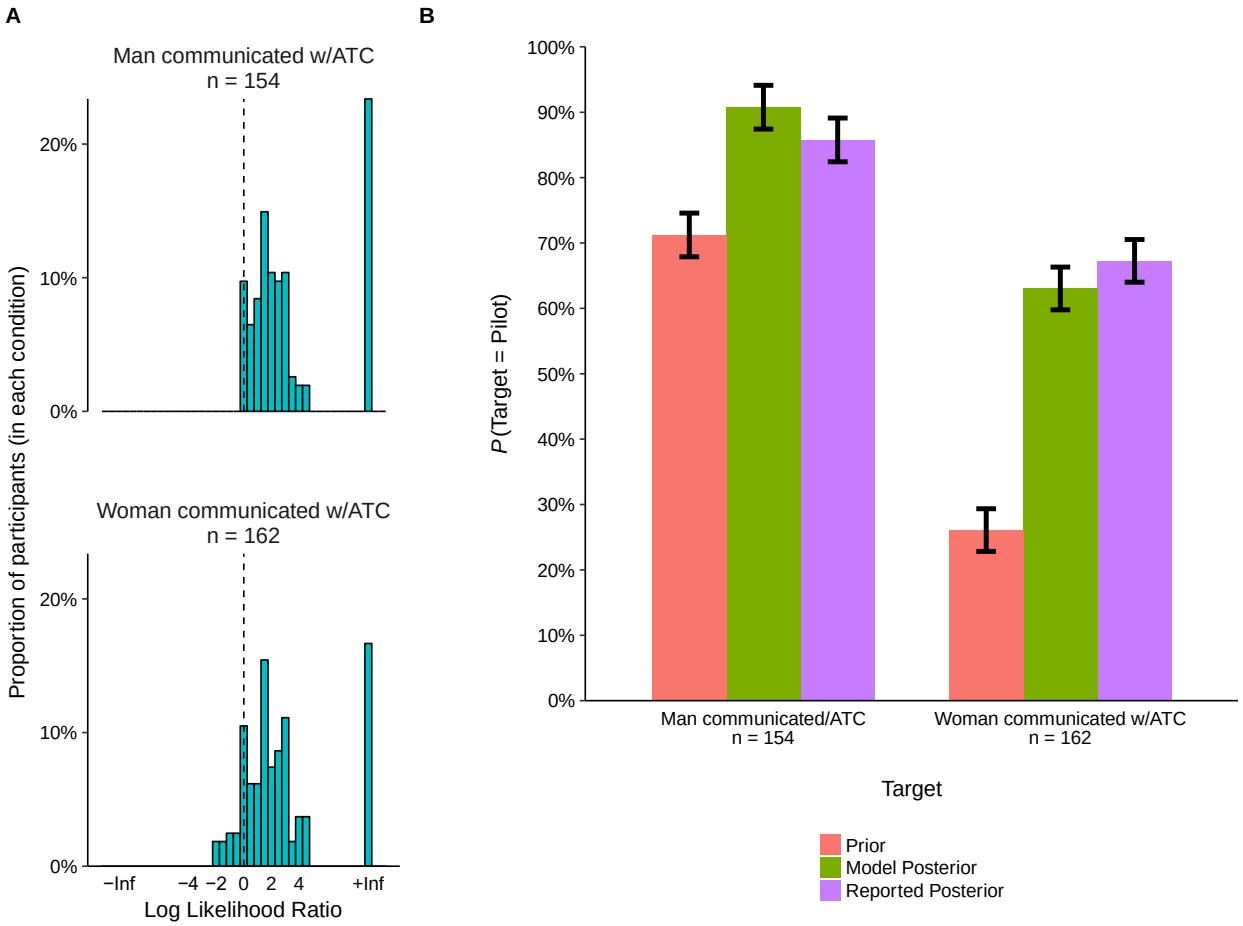


Fig. S12. Additional study that conceptually replicates Bayesian judgments: communicated with air traffic control (ATC) conditions. **A.** The correspondence between model and reported posteriors is present at the level of the individual participant. By subtracting each participant's model posterior 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, $B = 0.31$, $t(250) = 5.77$, $P < 0.0001$, $r = 0.34$.

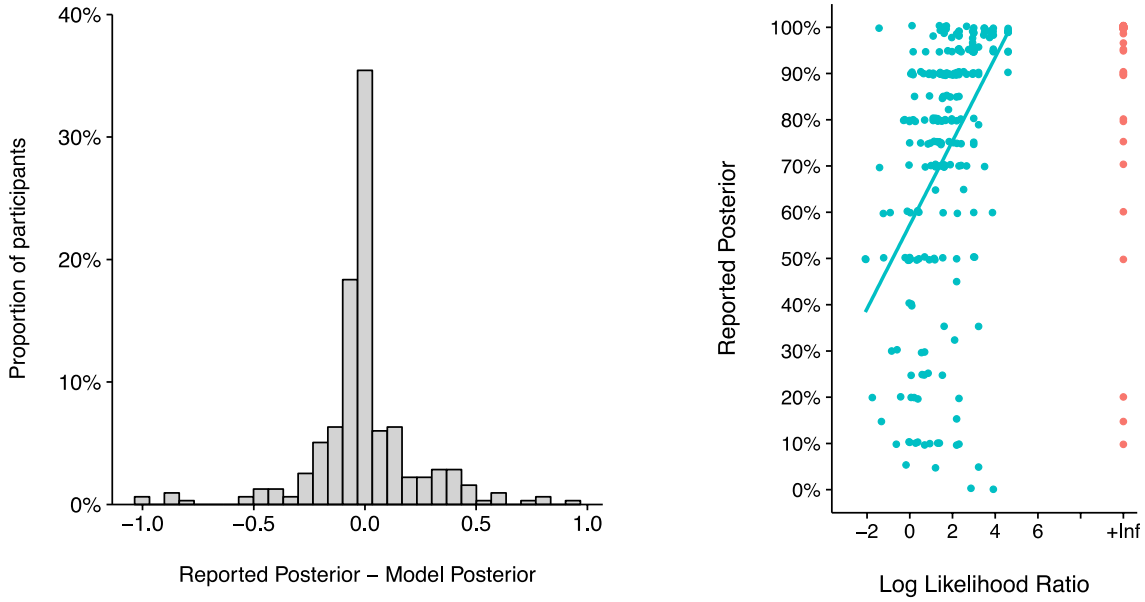


Fig. S13. Additional study that conceptually replicates Bayesian judgments: communicated with air traffic control (ATC) conditions. The statistical significance of the four critical comparisons is robust to the choice of adjustment factor when participants' probability judgments are logit transformed. 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.

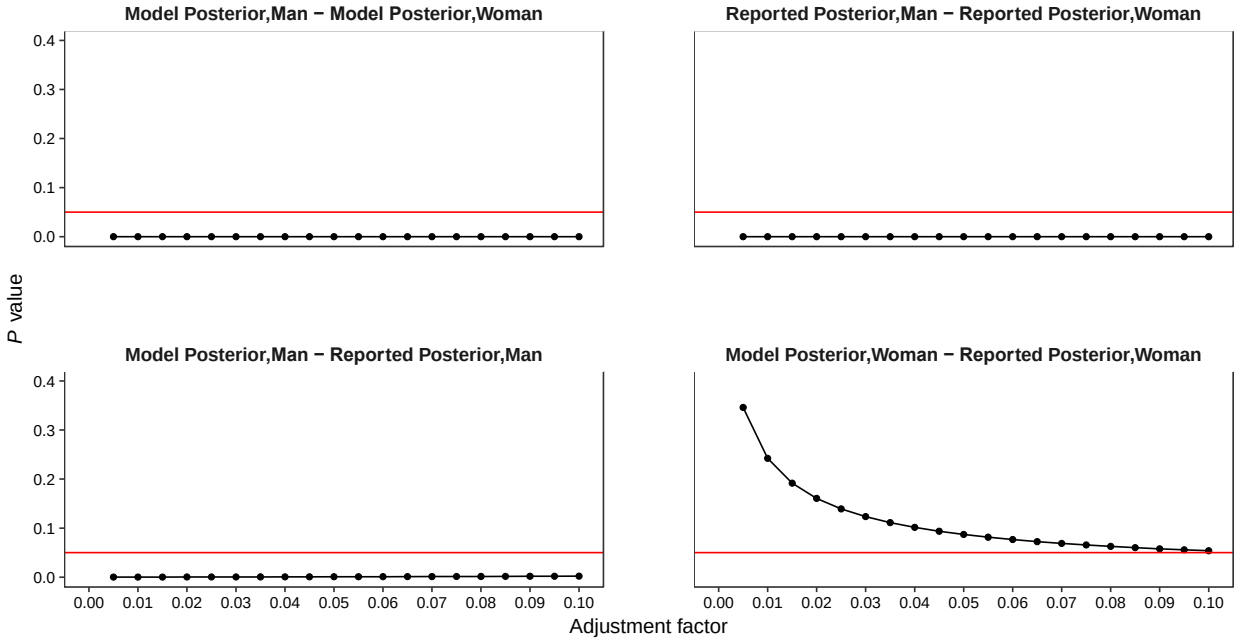


Fig. S14. Study 5. **A.** Distribution of likelihood ratios (log scaled) in each condition. **B.** Average judgments among participants in each condition. Priors indicate judgments before participants learned that the target had communicated with air traffic control. Model posteriors indicate judgments participants should make from a Bayesian perspective. Reported posteriors indicate judgments participants actually made. Error bars are 95% *CI*s.

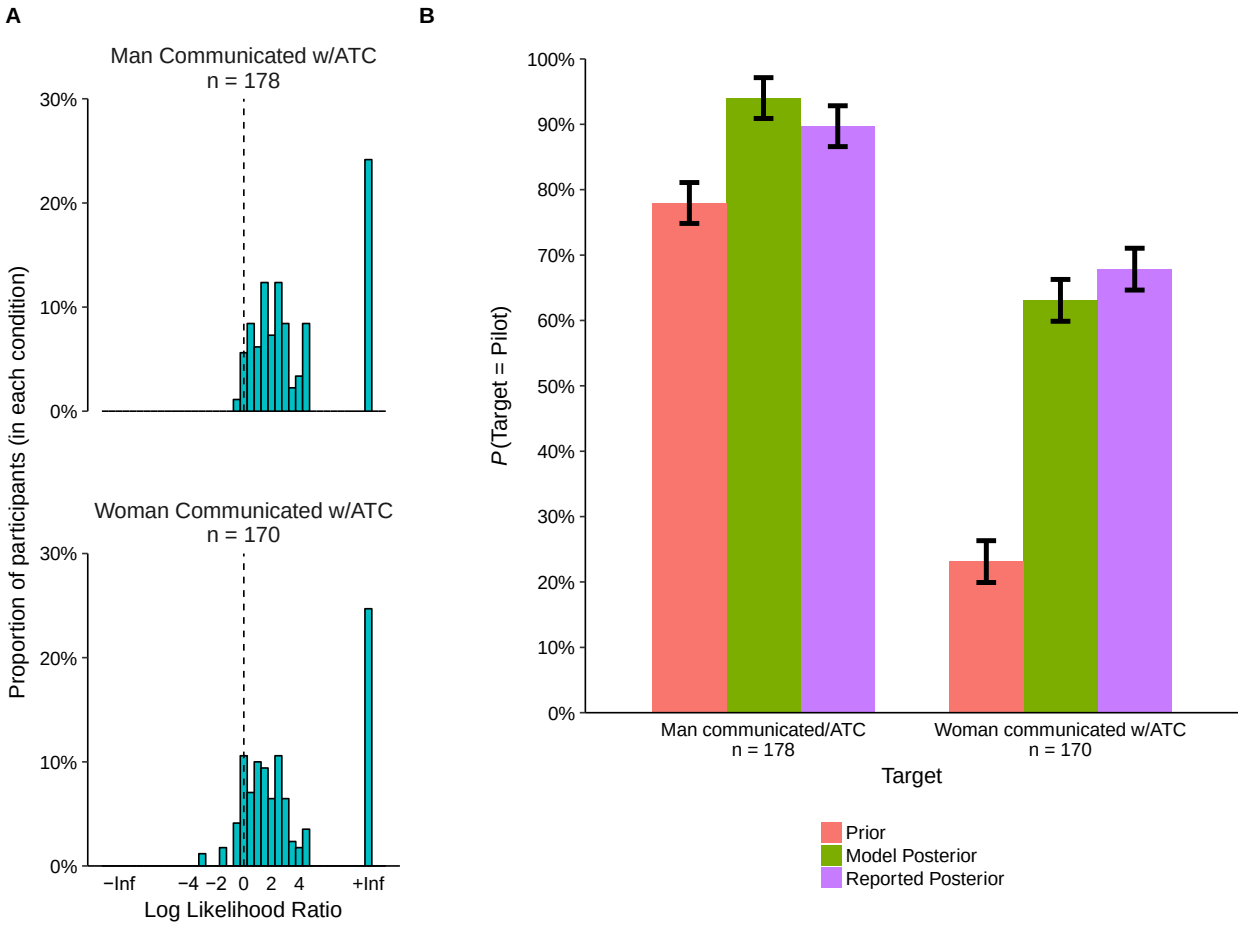
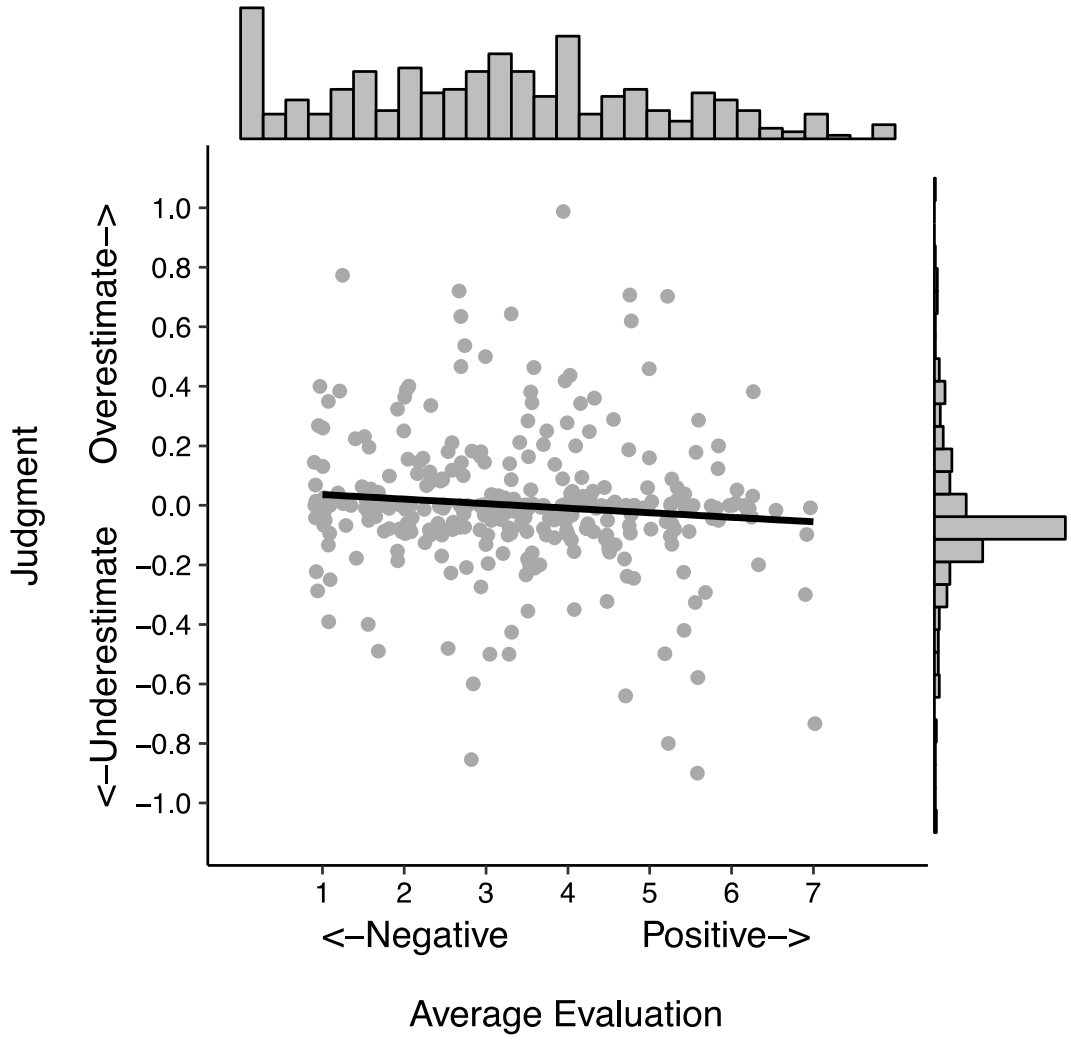


Table S5: Study 5. Proportion of participants who agreed that the woman is more likely to be a doctor, conditional on both the man and woman having performed surgery, that they're equally likely to be a doctor, or that the man is more likely to be a doctor (rows). Proportion of participants whose priors 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.

	Woman more likely to be pilot	*Equally likely to be pilot*	*Man more likely to be pilot*	
Woman more likely to be doctor	0%	0.29%	1.15%	1.44%
Equally likely to be doctor	1.15%	7.18%	70.69%	79.02%
Man more likely to be doctor	0.57%	0.29%	18.68%	19.54%
	1.72%	7.76%	90.52%	100%

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 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 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, $M_{\text{Model Posterior, Man}}$ = 85.3% vs. $M_{\text{Model Posterior, Woman}}$ = 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, $M_{\text{Reported Posterior, Man}}$ = 79.7% vs. $M_{\text{Reported Posterior, Woman}}$ = 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, $M_{\text{Model Posterior, Man}}$ = 85.3% vs. $M_{\text{Reported Posterior, Man}}$ = 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, $M_{\text{Model Posterior, Woman}}$ = 65.6%; vs. $M_{\text{Reported Posterior, Woman}}$ = 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

α = 0.89, $M_{composite}$ = 4.49, *SE* = 0.07, one-sample *t*(354) = 6.80, *P* < 0.0001, 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, η^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% *CI*s.

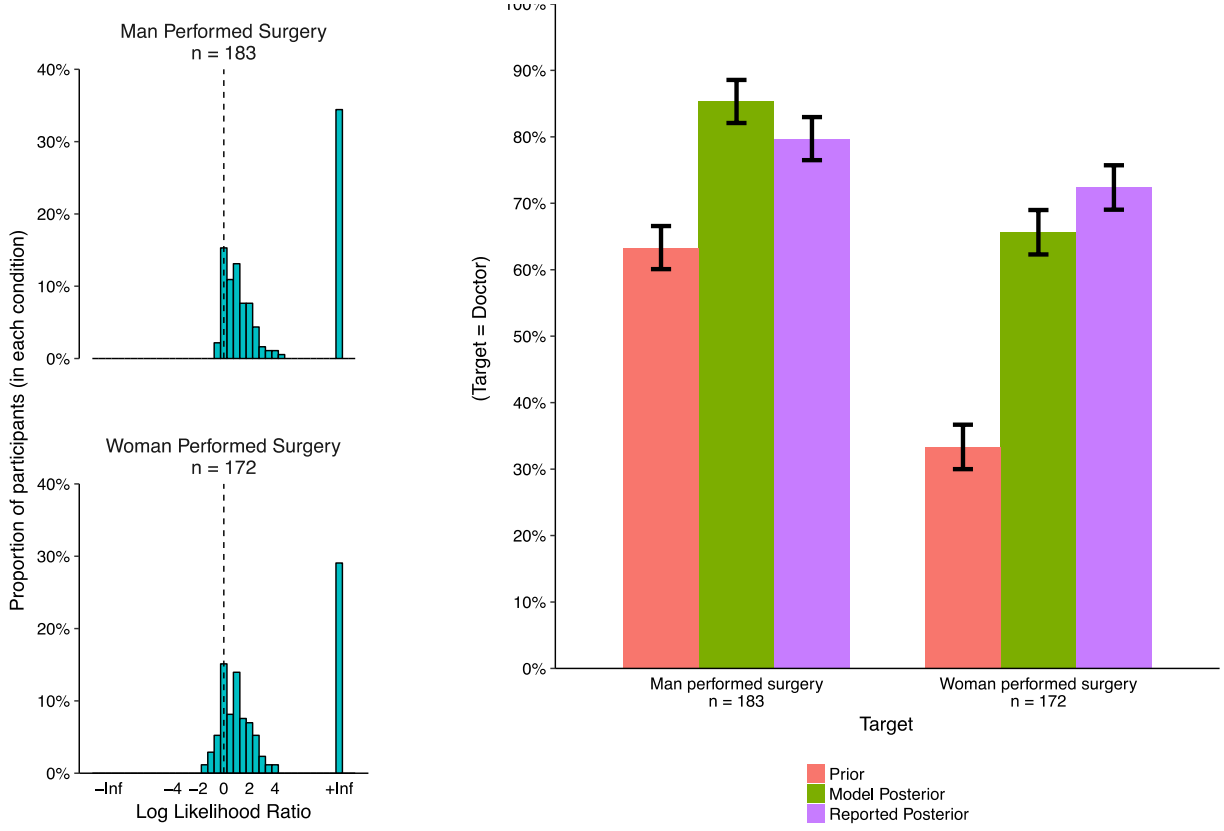


Fig. S17. Conceptual replication of Study 5. Reported posterior probabilities as a function of evaluations of Person X (average of four Likert-type items). Grey bands are *SE*s.

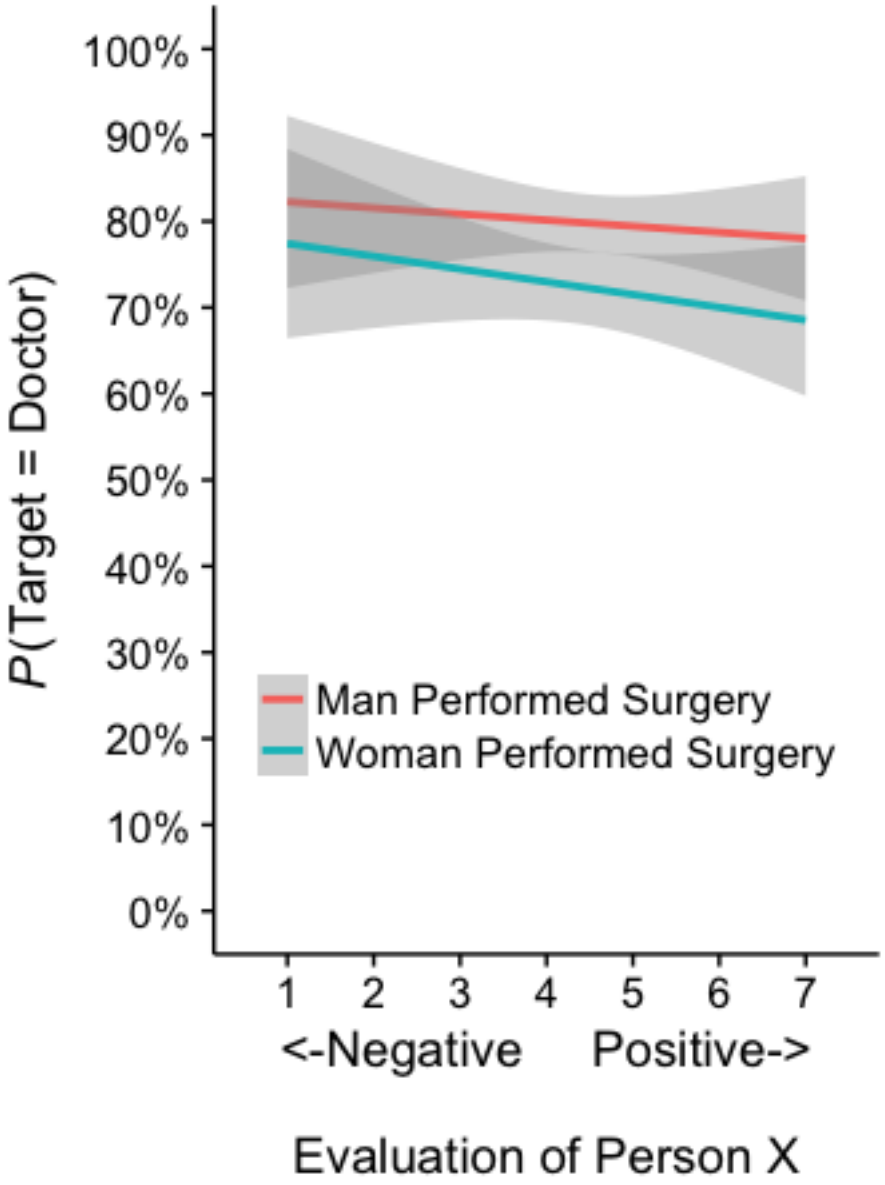


Fig. S18. Conceptual replication of Study 5. Scatterplot and line of best of fit showing the relationship between statistical accuracy on y-axis (model posterior subtracted from reported posterior) and evaluations of Person X on the x-axis (average of four Likert-type items). Distributions of each variable 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.