Supporting Information

Mahmoodi et al. 10.1073/pnas.1421692112

SI Methods

Experiment 1.

Participants. Participants were students at Aarhus University, Denmark ($n = 30$, mean age ± SD: 22.2 ± 2 y), the University of Tehran, Iran ($n = 30$, mean age ± SD: 24 ± 1.3 y), and Peking University, China ($n = 38$, mean age ± SD: 22.8 ± 2.2 y). We excluded two dyads (both from the Chinese cohort) because the respective dyad members were so similar in terms of their sensitivity that we could not reliably identify which one was less/more sensitive for the analysis shown in Fig. 2*A*. The exclusion of these participants did not affect the results of any of the other analyses. Members of each dyad knew each other beforehand. Participants received monetary compensation for their participation. No participant was recruited for more than one experiment. The local ethics committees approved each experiment, and written informed consent was obtained from all participants.

Display parameters and response mode. Participants sat at right angles to each other in a dark room; each with their own monitor and response device. The monitors [Denmark: resolution (pixels) = 1,680 × 1,050, Fujitsu Siemens AMILO SL 3220W, 22"; Iran: resolution = 1,280 × 1,024, Samsung SyncMaster 173p+, 17"; China: resolution = 1,600 × 1,200, ViewSonic p97f+, 19"] were linked to the same experimental computer via a video splitter and controlled by the Cogent toolbox (www.vislab.ucl.ac.uk/cogent.php/) for MATLAB (Mathworks). Viewing distance was ∼57 cm. We used a look-up table to linearize the output luminance of each monitor. The experimental code imposed the same display resolution (800 × 600) irrespective of the native resolution of the monitors. Background luminance was ∼62.5 Cdm^2. Within each session of the experiment, one participant responded with the keyboard and the other participant responded with the mouse. The participants swapped response devices after each session. A piece of thick black cardboard was used to occlude half of the display for each participant: the right half of the display for the participant responding with the keyboard, and the left half for the participant responding with the mouse. This was done to ensure that participants could make their individual responses in private (see below) despite receiving input from the same shared graphic card.

Stimulus. The stimulus shown in each interval consisted of six vertically oriented Gabor patches (SD of the Gaussian envelope: 0.458; spatial frequency: 1.5 cycles/°; luminance contrast: 10%) organized around an imaginary circle (radius: 8 visual degrees) at equal distances from each other. One of the patches had higher contrast (the target). The target contrast was obtained by randomly adding one of four values (1.5%, 3.5%, 7.0%, or 15%) to the 10% baseline contrast of the nontarget patches. The spatial location (which patch) and the temporal location (which interval) were randomized. A fixation cross was displayed at the center subtending 0.758° of visual angle in both the first and the second interval.

Procedure. Two participants (a dyad) sat in the same testing room, each viewing their own computer monitor. They performed a two-alternative forced-choice contrast discrimination task (Fig. 1*A*). On each trial (256 trials), dyad members viewed two consecutive intervals. First, a fixation cross appeared on the screen for a variable period, drawn uniformly from the range 500–1,000 ms. Next, participants viewed the two stimulus intervals, one of which contained the target. Each interval lasted 83 ms; the intervals were separated by a blank display lasting 1,000 ms. In each interval, there were six vertically oriented Gabor patches. In either the first or the second interval, one of the patches (the target) had a higher level of contrast.

After the two viewing intervals, dyad members privately indicated which interval they thought contained the target and how confident they felt about this decision (on a scale from 1 to 5 in discrete steps). A horizontal line appeared on the screen with a fixed midpoint. A vertical confidence marker was displayed on top of the midpoint. Color codes were used to denote the keyboard (blue) and the mouse (yellow) marker. Participants indicated which interval they thought contained the target and how confident they felt about this decision. Each participant made their private response by moving their marker to the left (first interval) or to the right (second interval) of the midpoint. The marker could be moved along the line by up to five steps on either side, with each step further away from the center indicating higher confidence. The keyboard participant navigated the marker by pressing N (move left) and M (move right) and confirmed their response by pressing B. The mouse participant navigated the marker by pressing the left (move left) and the right (move right) button and confirmed their response by pressing the scroll button. Note that at this stage, responses were made privately: each participant could only see their own marker.

Next, the participants' private responses (decision and confidence) were made public. If participants disagreed (i.e., if participants had privately selected different intervals), then a joint response was requested. The keyboard participant was nominated to make this response on odd trials and the mouse participant on even trials. The arbitrator had access to the responses of each dyad member, but no other communication was allowed. The nominated participant (the arbitrator) indicated the joint response by moving a white marker along the horizontal line; only the arbitrator could see the marker at this stage. After the arbitrator had indicated the joint response, all three markers (keyboard, mouse, and joint) were displayed together for 500 ms. Participants then received feedback about the accuracy of each decision; the feedback text read CORRECT or WRONG and was color-coded as detailed above. The vertical order of the individual feedback was randomized. The joint feedback always appeared at the center. If participants agreed (i.e., participants privately selected the same interval), then the joint decision was taken to be the same as the private decisions and participants continued directly to feedback. The feedback text remained on the screen until the participant using the keyboard initiated the next trial. Participants were not allowed to communicate verbally with each other; they were divided by a screen and given earphones to block out sound. The experimenter was present in the testing room throughout the experiment to ensure that the instructions were observed.

Experiment 2. Healthy Iranian adult males took part in experiment 2, which was conducted in Tehran University Iran ($n = 22$, mean age ± SD: 22 ± 2.5 y). This experiment differed from experiment 1 in one respect: in addition to the trial-specific feedback (see above), participants were presented with a running score of their own and their partner's accuracy at the end of each trial (i.e., the percentage of correct responses in all trials from the beginning of the experiment up until the current trial). This cumulative feedback ensured that participants did not have to rely on their memory for tracking their own or their partner's accuracy, thus testing the hypothesis that equality bias observed in experiment 1 might be due to limitations of memory.

Experiment 3. Participants were students at Aarhus University, Denmark ($n = 26$, mean age ± SD: 22.2 ± 2.4 y) and the University

of Tehran, Iran (n = 20, mean age ± SD: 23.4 ± 1.8 y). The experiment was divided into two sessions. In the first session, the two participants performed an isolated version of the psychophysical task described above. On each trial, participants viewed the stimulus intervals, made their individual private responses, and then received feedback about the accuracy of their own decision. There was no sharing of responses, no joint decision, and no shared feedback (128 trials). At the end of this session, we computed the sensitivity of each participant, and for the second session, increased the level of stimulus noise for the less-sensitive participant. In the second social session, participants performed the same task as in experiment 1. Participants were not informed about the noise assignment procedure until the very end of the experiment and were told that the first session served as practice. By means of this design, we created a significant difference in performance between the less and the more sensitive dyad members (Fig. S2), thus excluding that the equality bias observed in experiment 1 (see main text) was due to small or happenstance differences in performance.

Experiment 4. Participants were students at University College London, United Kingdom (n = 34, mean age ± SD: 24 ± 4.7 y). This experiment differed from experiment 1 in one respect: instead of rating their confidence, participants shared their choice uncertainty by postdecision wagering (PDW) (1). We wanted to test if monetary incentives would reduce or eradicate the equality bias by increasing the motivation for higher joint accuracy compared with experiment 1. In each trial, dyad members first made their individual choice and bet 0.20, 0.40, 0.60, 0.80, or 1.0£ on that decision. Choices and bets were then shared, and a joint choice and bet was then made by one of the two dyad members. Feedback was delivered adding up each participant's wins and losses from the two betting stages. An example scenario: (*i*) participant A initially bet 0.40£ on interval 1 and participant B bet 0.80£ on interval 2; (*ii*) the dyad then placed 0.60£ on interval 2; (*iii*) the correct answer was interval 2; and (*iv*) participant A won −0.40 + 0.80 = 0.40£, and participant B won 0.80 + 0.60 = 1.40£. Apart from the flat rate compensation for participating in the experiment, each participant received a bonus payment calculated by randomly selecting five trials and adding up the earnings in those trials.

Estimating Individual and Collective Performance. To quantify performance, separately for each decision maker (individual and dyad), we first plotted the proportion of trials in which the target was reported to be in the second interval against the contrast difference between the two intervals (Δc; i.e., the contrast level in the second interval minus the contrast level in the first interval at the target location; Fig. 1*C*). The resulting curves were then fit to a cumulative Gaussian function with the parameters bias, b, and variance, σ^2, using a probit regression model (*glmfit* function in MATLAB; Mathworks). A decision maker with bias b and variance σ^2 would have a psychometric function $P(\Delta c)$ given by

$$P(\Delta c) = H\left(\frac{\Delta c + b}{\sigma}\right), \qquad \text{[S1]}$$

where $H(z)$ is the cumulative normal function

$$H(z) \equiv \int_{-\infty}^{z} \frac{dt}{(2\pi)^{1/2}} \exp\left[-t^2/2\right]. \qquad \text{[S2]}$$

Given the above definitions for $P(\Delta c)$, we see that the decision variance is related to the maximum slope (denoted s) of the fitted psychometric curve at its point of inflection, via

$$s = \frac{1}{(2\pi\sigma^2)^{1/2}}. \qquad \text{[S3]}$$

A steeply rising curve has a large slope indicating small variance, which we interpreted as high sensitivity. We constructed psychometric functions for each dyad member and for the dyad. We defined collective benefit as the ratio of the dyad's sensitivity (s_{dyad}) to that of the more sensitive member of the dyad (i.e., the dyad member with higher sensitivity, $s_{\max}$). Thus, a value above 1 would indicate that the dyad outperformed its better member.

Computational Model. We modeled participants' beliefs about the reliability of their own and their partner's responses using a previously introduced computational model (2). In its original formulation, the model uses Bayesian reinforcement learning to track, trial by trial, the probability of obtaining a reward from taking a choice alternative. Broadly, it assumes that the choice outcome is generated by an underlying probability, p, and after having observed the outcome on a given trial, the model updates its estimate of p. Importantly, the model scales this update (the learning rate) according to the volatility of the environment. In a changing environment, p primarily reflects the history of recent trials. In a stable environment, updates should be smaller and p incorporates the history of more distant trials. We assumed that participants in our task used a mechanism similar to the one described above to track the probabilities of their own as well as their partner's response (choice alternatives) being correct (reward). In a sense, these probabilities reflect the reliability of the respective responses. We note that it has been shown previously that people indeed can track the reliability of more than one source of information (3). We obtained these (trial-by-trial) probabilities for each participant by fitting (maximum likelihood estimation) the model described above to the decisions about the target interval and their associated outcomes. We refer the reader to the original papers for mathematical details (3).

Estimating Social Influence. To estimate how a dyad member weighted their partner's decision relative to their own, we focused on those disagreement trials in which the dyad member made the joint decision on behalf of the dyad. We assumed that, on each disagreement trial i, the dyad member made a joint decision by (*i*) scaling the estimated probabilities of their own as well as their partner's decision being correct ($p_{s,i}$ and $p_{o,i}$, where s indicates self and o indicates other; see above), with the expressed levels of confidence ($c_{s,i}$ and $c_{o,i}$) and (*ii*) combining these scaled estimates into a decision criterion. Critically, to capture any biases, we introduced a free parameter, γ, which controlled the influence of the partner in the formation of the decision criterion (Eq. **1**).

Values of $\gamma > 1$ indicate that the dyad member is strongly influenced by their partner. In contrast, values of $\gamma < 1$ indicate that the dyad member is weakly influenced by their partner. We defined γ_{fit} as the γ value that maximized the fit (maximum likelihood estimation) between the model-derived and the empirically observed joint decisions. We defined γ_{opt} as the γ value that maximized the sensitivity of the model-derived joint decision. We used the discrepancy between these two values, $\gamma_{fit} - \gamma_{opt}$, to determine whether the dyad member underweighted ($\gamma_{fit} < \gamma_{opt}$) or overweighted ($\gamma_{fit} > \gamma_{opt}$) their partner's opinion relative to their own.

Under What Kind of Conditions Would the Decision Criterion Become Negative? As an example, *DC* would be negative in the following scenario:

i) The arbitrating participant expresses confidence 1 out of 5 in interval 2. That is to say $c_s = +1$.

ii) The partner expresses confidence 5 out of 5 in interval 1. That is to say $c_o = -5$.
iii) Given the history of the trials, the arbitrator believes that $p_{o,i} = 0.75$ and $p_{s,i} = 0.80$.
iv) The arbitrator adheres to equality bias i.e., $\gamma = 1$.

In this case, $DC = (0.80) \times (+1) + (1) \times (0.75) \times (-5) = 0.8 - 3.75 = -2.95$, the negative sign of which would indicate the joint decision will be the first interval.

Estimating the Combined Equality Bias. We quantified the dyad members' combined equality bias as $(\gamma_{opt,s_{min}} - \gamma_{fit,s_{min}}) - (\gamma_{opt,s_{max}} - \gamma_{fit,s_{max}})$. The equality bias is expected to drive the less and more sensitive dyad members in opposite directions, meaning that $(\gamma_{opt,s_{min}} - \gamma_{fit,s_{min}}) > 0$ and $(\gamma_{opt,s_{max}} - \gamma_{fit,s_{max}}) < 0$. We therefore reasoned that a dyad's combined equality bias would be best estimated by subtracting these two estimates from one another.

The Distribution of Data Points into Bins of the Factorial Design in Fig. 2C. The distribution of data points in the 2×2 bins of the factorial design used in Fig. 2*C* is important for the validity of the used ANOVA. There was a one-to-one correspondence for the factor dyad member as the labels better and worse were assigned to the two members within each dyad. Binning the trials according to decision accuracy resulted in (~70% correct) and (~30% wrong) split of the trials (average accuracy ± SD = 67 ± 6%) into correct and wrong individual decisions. As a results, we had around (256×0.7=) 160 correct and 90 incorrect trials to calculate P_{high} within each bin. Finally, the dependent variable P_{high} reflects a proportion within each bin.

Cross-Cultural Considerations. We used the same experimental procedure across all three locations, except that all instructions and task text were given in the local language. This procedure was chosen to eliminate any foreign language effects on behavior (4). We observed no difference in performance (sensitivity) between the three locations. In particular, independent *t* tests comparing Iranian and Danish participants [$t(58) = 1.6$, $P > 0.1$], Danish and Chinese participants [$t(66) = 1.03$, $P > 0.3$], and Iranian and Chinese participants [$t(66) = -0.6$, $P > 0.54$], were all nonsignificant.

Can Our Results Be Explained by Regression to the Mean? Regression to the mean may happen in situations in which (*i*) one variable is measured at two time points or (*ii*) two variables are measured at the same time point. In the first situation, if a variable is extreme (relative to the population distribution) on its first measurement, then it tends to be less extreme on its second measurement. Similarly, if a variable is extreme on its second measurement, it tends to have been less extreme on its first measurement. In the second situation, if one variable is measured to be extreme, and then another variable measured at the same time will tend to be less extreme, and vice versa.

Therefore, regression to the mean should be considered as a potential confound whenever data are split according to where data points fall along some distribution, as is (roughly) the case with our split of participants into less sensitive (s_{min}) and more sensitive (s_{max}) dyad members. However, we believe that there are two key reasons why regression to the mean cannot explain our finding that less sensitive dyad members tend to underweight their partner's opinion (i.e., optimality: $\gamma_{fit} - \gamma_{opt} < 0$), whereas more sensitive dyad members tend to overweight their partner's opinion (i.e., optimality: $\gamma_{fit} - \gamma_{opt} > 0$).

First, if sensitivity (s) and optimality ($\gamma_{fit} - \gamma_{opt}$) were two (uncorrelated) random variables, then, under regression to the mean, we would expect participants who were extreme in terms of sensitivity (e.g., very low or very high) to be less extreme in terms of optimality (i.e., $\gamma_{fit} - \gamma_{opt}$ approaching 0) and vice versa. However, we found, as indicated by the correlation shown in Fig. 3*A*, that participants who were extreme in terms of sensitivity (in particular, very low) also were extreme in terms of optimality (i.e., $\gamma_{fit} - \gamma_{opt}$ far from 0). Second, regression to the mean is more expected when data are split according to the distribution of values across the population. However, we did not categorize participants according to the distribution of sensitivity across participants but according to the dyadic relation between their own sensitivity and that of their partner. Therefore, a participant who was the less sensitive member of one dyad could have been the more sensitive member of another dyad, making it less plausible that our effects should be due to regression to the mean.

In addition to the above arguments, we also considered the possibility that dyad members' confidence showed some sort of regression the mean: the less sensitive dyad members might have increased their confidence over time, whereas the more sensitive dyad members decreased their confidence. To test this possibility, we divided our data into two time bins and calculated the average confidence in each time bin. Neither of the dyad members showed any significant change in confidence [Fig. S6; comparing the two time bins; more sensitive dyad members: $t(46) = 0.99$, $P = 0.32$; less sensitive dyad members: $t(46) = 0.04$, $P = 0.96$; paired *t* tests]. A 2 (dyad member: less vs. more sensitive) × 2 (time: first and second experiment session) ANOVA showed no significant main effect of time [Fig. S6; $F(1, 92) = 0.49$, $p > 0.48$]. There was also no significant interaction between dyad member and time [Fig. S6; $F(1, 92) = 0.4$, $p > 0.52$].

SI Results

Experiment 1.

Results for each country.

Collective benefit. In all three countries, collective benefit (CB), which we defined as the ratio of the sensitivity of the dyad to that of the more sensitive dyad member (i.e., s_{dyad}/s_{max}), was significantly higher when the more sensitive dyad member indicated the joint response than when the less sensitive dyad member indicated the joint response [Denmark: $t(14) = -2.51$, $P < 0.03$; Iran: $t(14) = -2.67$, $P < 0.02$; China: $t(16) = -2.07$, $P = 0.05$; paired *t* tests]. There were no differences in the collective benefit between the three countries. Independent *t* tests comparing Iranian and Danish participants [$t(28) = -1.28$; $P > 0.21$], Danish and Chinese participants [$t(30) = -0.84$, $P > 0.4$], and Iranian and Chinese participants [$t(30) = 0.49$, $P > 0.62$] did not show any significant difference.

Following the better dyad member. The proportion of disagreement trials—irrespective of who was the arbitrator—in which the dyad decision was the same as that of the more sensitive dyad member was only significantly higher than 50% when the data were collapsed across the three countries. However, the data from the three countries showed the same pattern as presented in the main text [Denmark: $t(14) = 2.93$, $P < 0.02$; Iran: $t(14) = 1.46$, $P = 0.16$; China: $t(16) = 1.44$; $P = 0.16$; one-sample *t* tests]. Independent *t* tests comparing Iranian and Danish participants [$t(28) = 0.25$, $P > 0.8$], Danish and Chinese participants [$t(30) = -0.33$, $P > 0.74$], and Iranian and Chinese participants [$t(30) = -0.45$, $P > 0.65$] indicated no significant difference.

Average confidence. The confidence of the more sensitive dyad members was only significantly higher than the less sensitive dyad members when we collapsed the data across the three countries. However, the data from the three countries showed the same pattern as presented in the main text [Denmark: $t(14) = 1.46$, $P = 0.65$; Iran: $t(14) = 2.18$, $P = 0.04$; China: $t(16) = 1.29$, $P = 0.21$; paired *t* tests].

Average confidence in correct and error trials. The difference in confidence on correct and error trials between the less and the more sensitive dyad members only reached statistical significance when we collapsed the data across the three countries. However, the data from all three countries showed the same pattern as

presented in the main text (Fig. 1*C*). The less sensitive dyad members were more likely to make a high confidence error than their more sensitive partners [Denmark: $t(14) = -1.19$, $P = 0.25$; Iran: $t(14) = -1.18$, $P = 0.25$; China: $t(16) = -1.31$, $P = 0.2$; paired t tests]. Conversely, the more sensitive dyad members were more likely to make a high confidence correct decision than their less sensitive partners [Denmark: $t(14) = 1.5$; $P = 0.15$; Iran: $t(14) = 2.05$, $P = 0.05$; China: $t(16) = 0.8$, $P = 0.43$; paired t tests].

Selecting partner's decision. The proportion of trials in which the less sensitive dyad member selected their partner's decision as the joint decision was as follows: Denmark, 57 ± 17%; Iran, 41 ± 19%; China, 51 ± 20% (mean ± SD). The proportion of trials in which the more sensitive dyad member selected their partner's decision as the joint decision was as follows: Denmark, 45 ± 5%; Iran, 36 ± 11%; China, 44 ± 14% (mean ± SD). The difference in these proportions between the less and the more sensitive dyad member only reached statistical significance when we collapsed the data across the three countries. However, the data from all three countries showed the same trend as presented in the main text, with the less sensitive dyad members selecting their partner's decision as the joint decision more often [Denmark: $t(14) = -2.61$, $P = 0.02$; Iran: $t(14) = -0.75$, $P = 0.46$; China: $t(16) = -1.12$, $P = 0.27$; paired t tests].

Additional analyses.

Joint decision accuracy when confirming self vs. other. To provide a clearer description of the join decision making process, we separately calculated the proportion of trials in which participants correctly selected their own or their partner's decision as the joint decision. The more sensitive dyad members correctly selected their own decision more often than the less sensitive dyad members [69 ± 10% vs. 46 ± 13% (mean ± SD), $t(47) = 8.23$, $P < 10^{-10}$, paired t test]. In other words, the more sensitive dyad members appeared to be better judges of when to follow their own decision. This effect was found in all three countries [Denmark: $t(14) = 3.7$, $P = 0.002$; Iran: $t(14) = 5.82$, $P = 10^{-4}$; China: $t(16) = 4.99$, $P = 10^{-3}$; paired t tests]. In contrast, the less sensitive dyad members correctly selected their partner's decision more often than the more sensitive dyad members [57 ± 13% vs. 67 ± 12% (mean ± SD), $t(47) = -3.57$, $P = 10^{-3}$, paired t test]. In other words, the less sensitive dyad members appeared to be better judges of when to follow their partner's decision. This effect was found in all three countries [Denmark: $t(14) = -2.37$, $P = 0.03$; Iran: $t(14) = -1.48$, $P = 0.16$; China: $t(16) = -2.92$, $P = 0.009$; paired t tests].

Confirming self vs. other for correct and wrong joint decisions. To get a better picture of interactive behavior during joint decisions, we separated the data based on correct and wrong joint decisions and then calculated the proportion of trials within each set that the more and less sensitive partners confirmed their partners' decisions (Fig. S5). When the more sensitive dyad members made a correct joint decision, they were significantly less likely (than chance level) to have confirmed their inferior partner [one-sample t test comparing to 50%; $t(46) = -7.36$, $P < 10^{-8}$; Fig. S5, *Left*, black bar]. This result was also replicated across countries [for Danish participants; $t(14) = -5.95$, $P < 10^{-4}$; for Iranian participants $t(14) = -4.21$, $P < 10^{-3}$; for Chinese participants $t(16) = -4.19$, $P < 10^{-3}$; one-sample t test]. Conversely, in the trials where the more sensitive dyad members made a wrong joint decision, they were at chance between confirming themselves vs. partner [paired t test; $t(46) = -0.04$, $P > 0.96$; Fig. S5, *Left*, white bar]. The opposite pattern was observed for the less sensitive dyad members. When making a correct joint decision for group, the inferior partner was at chance for confirming self vs. other [one-sample t test; $t(46) = 1.53$, $P > 0.13$; Fig. S5, *Right*, black bar]. When making a wrong joint decision, the less sensitive dyad members were significantly less likely to have confirmed their more competent partner [one-sample t test; $t(14) = -4.45$, $P < 10^{-4}$; Fig. S5, *Right*, white bar]. This pattern was observed in all countries and was significant in Iran and China [for Danish participants; $t(14) = -0.28$, $P = 0.77$; for Iranian participants $t(14) = -5$, $P < 10^{-3}$; for Chinese participants $t(16) = -3.16$, $P < 0.006$; one-sample t test].

To see what circumstances led high performers to switch inappropriately to the low performer's initial decision, we looked at the trials in which more sensitive dyad members wrongly confirmed their partners. In 95 ± 8% (mean ± SD) of these trials, the less sensitive dyad members were as confident as or more confident than more sensitive dyad members.

Goodness of the model fit. Under γ_{fit}, the model predicted the participants' choice in about 83 ± 7% of all disagreement trials (Denmark: 85 ± 7%; Iran: 83 ± 7%; China: 82 ± 7%, mean ± SD). The model thus provided a good fit to the empirical data. There were no between-country differences in the proportion of disagreement trials explained. Independent t tests comparing Iranian and Danish participants [$t(58) = -0.76$, $P > 0.44$], Danish and Chinese participants [$t(62) = 0.93$, $P > 0.35$], and Iranian and Chinese participants [$t(62) = -0.93$, $P > 0.35$] were all nonsignificant.

Under- and overweighting. In all three countries, the less sensitive dyad members underweighted (i.e., $\gamma_{fit} < \gamma_{opt}$) the opinion of their more sensitive partners [Denmark: $t(14) = -2.17$, $P < 0.04$; Iran: $t(14) = -3.77$, $P < 0.002$; China: $t(16) = -4.55$, $P < 10^{-3}$; paired t tests]. There were no differences in the degree of underweighting (i.e., $\gamma_{opt} - \gamma_{fit}$) between the three countries. Independent t tests comparing Iranian and Danish participants [$t(28) = -0.26$, $P > 0.79$], Danish and Chinese participants [$t(30) = -0.61$, $P > 0.54$], and Iranian and Chinese participants [$t(30) = -0.41$, $P > 0.67$] were all nonsignificant. In contrast, we only found a significant effect of the more sensitive dyad members overweighting (i.e., $\gamma_{fit} > \gamma_{opt}$) their partner's opinion when we collapsed the data across the three countries. However, the data from all three countries showed the same pattern as presented in the main text [Denmark: $t(14) = 1.41$, $P = 0.17$; Iran: $t(14) = 1.55$, $P = 0.14$; China: $t(16) = 1.04$, $P = 0.31$; paired t tests]. There were no differences in the degree of overweighting (i.e., $\gamma_{opt} - \gamma_{fit}$) between the three countries. Independent t tests comparing Iranian and Danish participants [($t(28) = 0.23$; $P > 0.81$], Danish and Chinese participants [$t(30) = -0.2$, $P > 0.98$], and Iranian and Chinese participants [$t(30) = 0.22$, $P > 0.82$] were all nonsignificant.

Relating the model-based analysis to the behavioral data. We used the optimal weight (γ_{opt}) of each participant to calculate the number of arbitration trials in which the model recommended that the participant should select their partner's decision as the joint decision. The less sensitive dyad members selected their partner's decision significantly less often than recommended by the optimal model, with the data showing the same effect in all three countries [Denmark: $t(14) = -1.43$, $P = 0.14$; Iran: $t(14) = -5.97$, $P < 10^{-3}$; China: $t(16) = -2.69$, $P = 0.01$; paired t tests]. Conversely, the more sensitive dyad members selected their partner's decision significantly more often than recommended by the optimal model, with the data showing the same trend in all three countries [Denmark: $t(14) = 2.68$, $P = 0.01$; Iran: $t(14) = 1.6$, $P = 0.14$; China: $t(16) = 2.08$, $P = 0.05$].

Insight into own relative to partner's reliability. We quantified how good participants were at weighting their partner's opinion relative to their own as the absolute difference between the optimal and the fitted weight, $|\gamma_{fit} - \gamma_{opt}|$. This quantity was larger for the less sensitive dyad members (they were worse), with the data showing the same pattern in all three countries [Denmark: $t(14) = 3.64$, $P < 0.003$; Iran: $t(14) = 2.89$, $P < 0.02$; China: $t(16) = 1.76$, $P = 0.09$; paired t tests].

Correlation between sensitivity and deviation from optimal weighting. The negative correlation between participants' sensitivity and their deviation from the optimal weight (i.e., $|\gamma_{fit} - \gamma_{opt}|$) varied in

strength between the three countries (Denmark: $r = -0.35$, $P = 0.05$, $R^2 = 0.12$; Iran: $r = -0.31$, $P = 0.08$, $R^2 = 0.1$; China: $r = -0.25$, $P = 0.15$, $R^2 = 0.13$; Pearson's r) but showed the same pattern.

Equality bias. One prediction of the equality bias was that the fitted weights should be closer to 1 than the optimal weights (i.e., $|1 - \gamma_{fit}| < |1 - \gamma_{opt}|$). This prediction was significant in Denmark [$t(29) = -3.28$, $P < 0.003$, paired t test] and China [$t(29) = -3.28$, $P < 0.003$, paired t test] and showed a similar trend in Iran [$t(29) = -0.95$, $P = 0.31$, paired t test].

Experiment 2.

Additional results.

Goodness of the model fit. Under γ_{fit}, the model predicted the participants' choice in about 83 ± 8% of all disagreement trials.

Experiment 3.

Results for each country. The results within each country were consistent with the overall pattern reported in the main text. The less sensitive dyad members underweighted (i.e., $\gamma_{fit} < \gamma_{opt}$) the opinion of their more sensitive partners [Fig. S3, black bars; Denmark: $t(14) = -2.19$, $P = 0.04$; Iran: $t(14) = -2.55$, $P = 0.03$; paired t tests]. Conversely, the more sensitive dyad members showed a similar trend for overweighting (i.e., $\gamma_{fit} > \gamma_{opt}$) the opinion of their less sensitive partners [Fig. S3, white bars; Denmark: $t(14) = 1.07$, $P = 0.3$; Iran: $t(14) = 2.07$, $P = 0.06$].

Additional results.

Goodness of the model fit. Under γ_{fit}, the model predicted the participants' choice in about 82 ± 6% of all disagreement trials (Denmark: 83 ± 7%; Iran: 81 ± 5%, mean ± SD). The model thus provided a good fit to the empirical data. There were no between-country differences in the proportion of disagreement trials explained. Independent t tests comparing Iranian and Danish participants [$t(44) = 1.09$, $P > 0.27$] were nonsignificant.

Experiment 4.

Additional results.

Goodness of the model fit. Under γ_{fit}, the model predicted the participants' choice in about 84 ± 6% (mean ± SD) of all disagreement trials. The model thus provided a good fit to the empirical data.

1. Persaud N, McLeod P, Cowey A (2007) Post-decision wagering objectively measures awareness. *Nat Neurosci* 10(2):257–261.
2. Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS (2007) Learning the value of information in an uncertain world. *Nat Neurosci* 10(9):1214–1221.
3. Behrens TE, Hunt LT, Woolrich MW, Rushworth MF (2008) Associative learning of social value. *Nature* 456(7219):245–249.
4. Keysar B, Hayakawa SL, An SG (2012) The foreign-language effect: Thinking in a foreign tongue reduces decision biases. *Psychol Sci* 23(6):661–668.

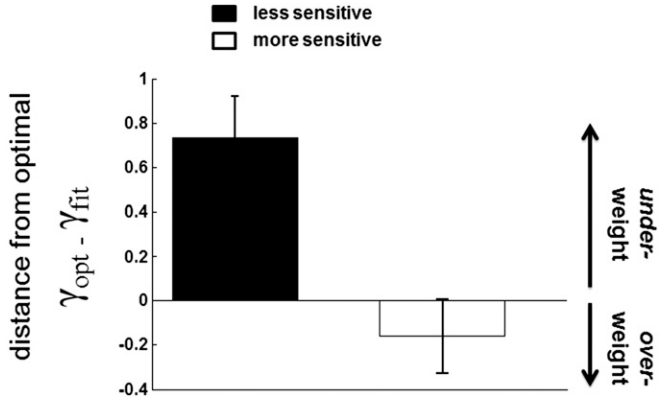

Fig. S1. Experiment 2. Deviation from the optimal weight ($\gamma_{opt} - \gamma_{fit}$) is plotted for the less sensitive (black bars, s_{min}) and the more sensitive (white bars, s_{max}) dyad members. Error bars are 1 SEM.

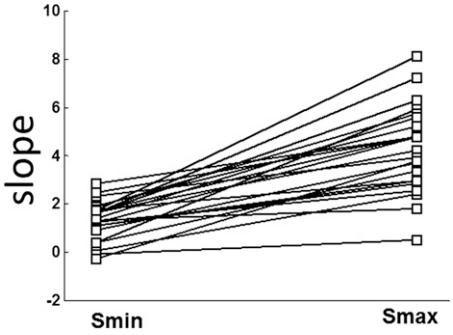

Fig. S2. Experiment 3. The slope of the psychometric function for the less sensitive, s_{min}, and the more sensitive, s_{max}, dyad members. Each line corresponds to a dyad. As intended, covert insertion of noise to the stimuli presented to one participant created a substantial difference between slopes of the two dyad members.

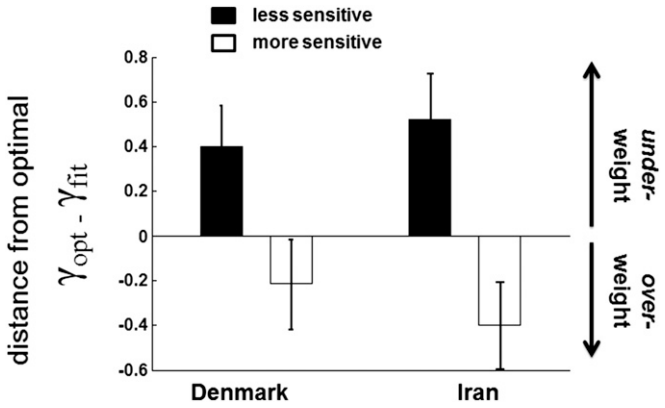

Fig. S3. Experiment 3. Deviation from the optimal weight ($\gamma_{opt} - \gamma_{fit}$) is plotted for the less sensitive (black bars, s_{min}) and the more sensitive (white bars, s_{max}) dyad members. Error bars are 1 SEM.

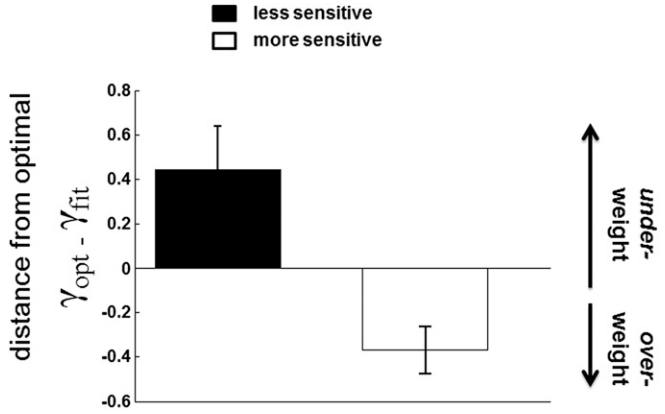

Fig. S4. Experiment 4. Deviation from the optimal weight ($\gamma_{opt} - \gamma_{fit}$) is plotted for the less sensitive (black bars, s_{min}) and the more sensitive (white bars, s_{max}) dyad members. Error bars are 1 SEM.

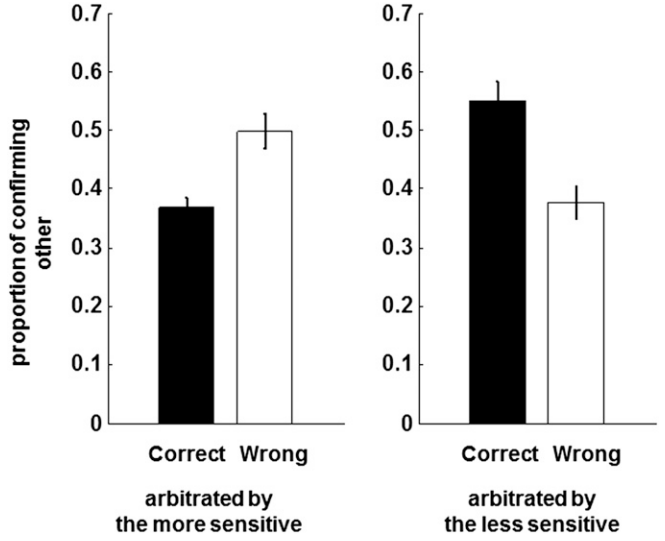

Fig. S5. Experiment 1. Proportion of trials in which the arbitrator confirmed his partner calculated separately for correct or wrong joint decisions. (*Left*) More sensitive dyad members. (*Right*) Less sensitive dyad members.

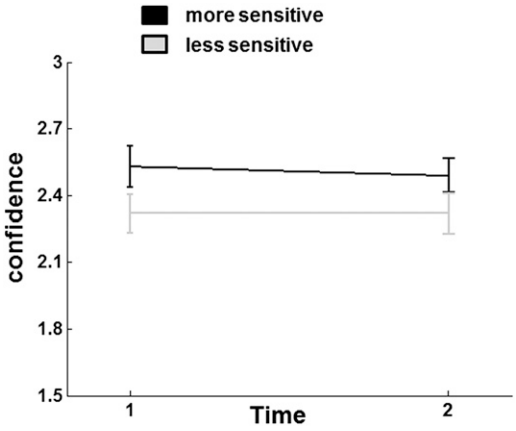

Fig. S6. Experiment 1. Change of confidence through time (first vs. second session) is plotted for the more (black line) and less (gray line) sensitive dyad members separately. Error bars are 1 SEM.