Learning rules of engagement for social exchange within and between groups

Rojek-Giffin, M.; Lebreton, M.; Daunizeau, J.; Fariña, A.; Gross, J.; Dreu, C.K.W. de

PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES
ECONOMIC SCIENCES

Learning rules of engagement for social exchange within and between groups

Michael Rojek-Giffin[a,1], Maël Lebreton[b,c,1,2], Jean Daunizeau[d,e], Andrea Fariña[a], Jörg Gross[a,f], and Carsten K. W. De Dreu[a,g,2]

Edited by Eric-Jan Wagenmakers, Universiteit van Amsterdam, Amsterdam, Netherlands; received November 7, 2022; accepted March 29, 2023 by Editorial Board Member Adrian E. Raftery

Globalizing economies and long-distance trade rely on individuals from different cultural groups to negotiate agreement on what to give and take. In such settings, individuals often lack insight into what interaction partners deem fair and appropriate, potentially seeding misunderstandings, frustration, and conflict. Here, we examine how individuals decipher distinct rules of engagement and adapt their behavior to reach agreements with partners from other cultural groups. Modeling individuals as Bayesian learners with inequality aversion reveals that individuals, in repeated ultimatum bargaining with responders sampled from different groups, can be more generous than needed. While this allows them to reach agreements, it also gives rise to biased beliefs about what is required to reach agreement with members from distinct groups. Preregistered behavioral (N = 420) and neuroimaging experiments (N = 49) support model predictions: Seeking equitable agreements can lead to overly generous behavior toward partners from different groups alongside incorrect beliefs about prevailing norms of what is appropriate in groups and cultures other than one's own.

Bayesian modeling | social neuroscience | bargaining | cooperation | beliefs

Significance

Thwarted intergroup relations and conflict often result from misunderstanding what outsiders find fair and appropriate, rather than unfair and offensive. Using computational modeling and behavioral experiments, we find when and how such misunderstandings can be avoided, allowing humans to interact with culturally distinct others in a constructive and mutually beneficial way. At the same time, in the process of learning "what it takes" to agree and avoid conflict, humans can come to believe that culturally distinct others need more than is actually the case. Findings can help to promote intergroup cooperation and trade, and to avoid intergroup conflict and polarization.

Many social interactions are governed by rules. From tipping in a restaurant to greeting rituals and extending and returning favors, humans tacitly develop and use implicit rules that enable them to negotiate transactions, sustain cooperation, and avoid coordination failures and conflict (1, 2). These implicit rules evolve over time and can become socially shared norms akin to a "secret code that is written nowhere, known by none, and understood by all" (3). However, groups can differ markedly in the norms they develop and, accordingly, the expectations they have about what behavior is deemed acceptable (4).

Implicit rules of engagement that groups develop locally, alongside the expectations of "what it takes to agree", can pose an important problem for intergroup interactions (5–7). Operating on rules of engagement one has learned within one cultural context can lead to surprise and frustration in interaction partners socialized with distinctly different ideas about what is fair and appropriate. Such violations of expectations can give rise to cooperation failures, social rejection, and conflict (8). To give a stylized example, consider an individual offering some share of a resource to their interaction partner, who can decide to either accept or reject the offer. Because rejection earns both proposer and responder nothing (9), proposers want to offer a share that meets the responder's acceptance threshold, while at the same time not offering more than needed for the responder to agree. And while proposers may have some intuition about their responder's acceptance threshold, for example based on what they themselves would deem fair and appropriate (1, 6, 10, 11), such intuition may be wrong especially when responders are from distinct cultural groups with very different fairness considerations (6, 10). Accordingly, to develop some joint course of action with people from different groups, humans need to accurately decipher their partners' implicit expectations and adapt their behavior. Failure to do so may thwart cross-boundary cooperation and even lead to intergroup conflict.

At present, we poorly understand whether and how humans can learn others' implicit rules of engagement, and with what consequences for social perception and cross-group cooperation. Whereas people often rely on group-based stereotypes and beliefs (8, 12), how these stereotypes and beliefs about another group's rules of engagement develop ex nihilo remains to be identified (13). Here, we seek answers to these questions and examine whether and how the individual's own conceptions of what is fair and appropriate—their social preferences—shape their understanding of unknown others' rules of engagement. We developed a computational model of individuals as Bayesian learners with inequality aversion (10, 14) who engage in ultimatum bargaining as proposers with unknown others as responders. Responders are sampled from different groups, each with a different but unknown rule of engagement (i.e., acceptance thresholds). We assumed that proposals are conditioned by the emotional aversion of giving others an unfair share, and by the

Author contributions: M.R.-G., M.L., J.G., and C.K.W.D.D. designed research; M.R.-G., M.L., A.F., and C.K.W.D.D. performed research; M.R.-G., M.L., and J.D. contributed new reagents/analytic tools; M.R.-G., M.L., A.F., and J.G. analyzed data; and M.R.-G., M.L., J.G., and C.K.W.D.D. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. E.-J.W. is a guest editor invited by the Editorial Board.

[1]M.R-G and M.L. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: mael.lebreton@psemail.eu or c.k.w.de.dreu@fsw.leidenuniv.nl.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2218443120/-/DCSupplemental.

Published May 1, 2023.

expectation that others reject something they deem unfair (2). Our model reveals how individuals, over time and through repeated exposure to accept/reject decisions by responders from different groups, gradually learn and adapt to the implicit rules and expectations embedded in these different groups. Crucially, however, we show also that inequality aversion can lead individuals to behave more generously than needed, inducing biased beliefs about "what it takes to agree" and leading individuals to continue to offer more than needed to secure agreement. Behavioral and neuroimaging experiments validated several key predictions of the model, providing the foundations of a neurocomputational account of cooperation and agreement within and between culturally distinct groups.

Results

For our analysis, we first created three pools of responders that differed in what ultimatum offers they would accept or reject. In a second step, we developed a learning model of proposers that update behavior based on observing responders' decisions and, in a third step, tested model predictions in several behavioral experiments (*SI Appendix*, Fig. S1). Results not only reveal to what extent individuals learn and adapt to interaction partners from different responder groups, but also how learning shaped the individuals' stereotypical beliefs about these different groups. Finally, neuroimaging experiments validated core assumptions of our computational model of humans as Bayesian learners with social preferences.

Creating and Modeling Different Responder Groups. To create responder groups that differed in what ultimatum offers they would accept or reject, we asked 210 participants, as responders, whether they would accept or reject a range of possible offers from proposers out of an endowment of 20 monetary units (MU). We manipulated the participant's starting endowments (from 0 to 20 MU, see *Materials and Methods*). Consistent with other experiments (9), responders endowed with 0 MU were most likely to accept offers that gave them at least 50%. As expected, however, responders endowed with 10 or 20 MU more likely accepted offers that gave them (far) less than 50% (Fig. 1*A*).

We fitted responders' decisions with logistic choice functions that capture the probability of each offer being accepted in each group with two parameters: an intercept (θ_1) and a slope (θ_2; Fig. 1*B* and *SI Appendix*, *Section I.2*). These functions formally

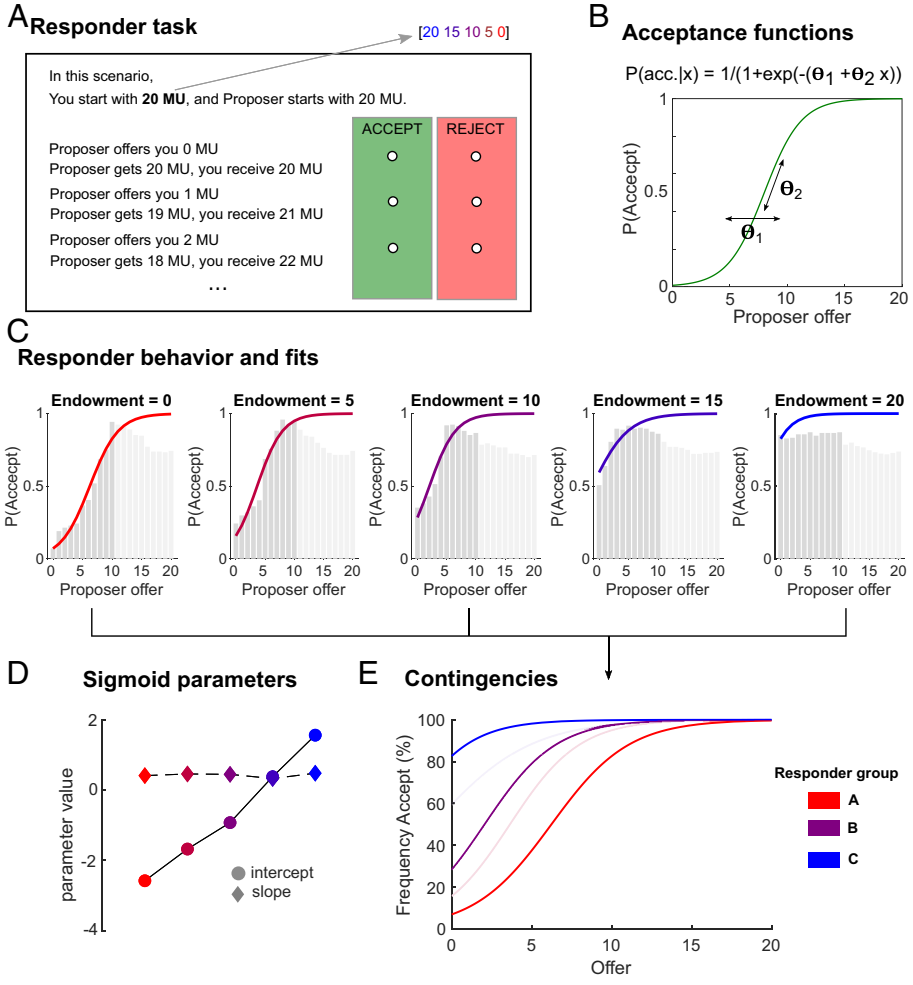

Fig. 1. Creating responder groups. (*A*) Participants provided accept/reject responses for all possible offers for five different starting endowments. (*B*) General acceptance functions as sigmoids with intercept θ_1 and slope θ_2. (*C*) Group-specific acceptance thresholds, estimated from responder data. Gray histograms represent acceptance frequencies of responders from the three different responder groups created in the laboratory to any possible offer between 0 and 20. Responder groups differed in their acceptance threshold, akin to diverging fairness norms. Colored lines represent acceptance probabilities as sigmoid functions fitted to the data to characterize the different responder groups' acceptance function. (*D*) Group-specific acceptance function parameters. (*E*) Responder groups selected to investigate proposer's behavior. The feedback to proposers in Experiments 2 to 5 was derived from the response functions for responder endowments of 0, 10, and 20.

describe the underlying acceptance probability of the three different responder groups and delineate what proposers' offers need to converge upon when repeatedly interacting with responders sampled from these different groups (Fig. 1*C*). Supporting that the experimental manipulations created different responder groups that mostly affected responders' acceptance threshold, acceptance functions were characterized by very similar slopes across the different groups, but intercepts varied linearly as a function of the endowment asymmetry (R^2 = 0.986; Fig. 1 *D* and *E*). These functions are agnostic to the responders' reasons for accepting or rejecting an offer (e.g., inequality aversion, reputational concerns), and mimic a group-specific behavioral norm about how generally acceptable different offers are.

Modeling Proposers' Behavior. To model how proposers may learn different responder functions and adapt their behavior to the different acceptance thresholds, recall that i) proposers should make offers that are large enough to surpass what the responder deems acceptable, yet ii) offering more than necessary to secure agreement reduces proposer's share of the pie (2, 7, 9). Agents fully informed about the different groups' acceptance functions can compute the expected gain of each possible offer by multiplying the probability of an offer being accepted with the monetary outcome (i.e., the endowment minus the offer) and select the offer that maximizes this expected payoff (Fig. 2*A*). Yet, when acceptance functions are unknown, agents must rely on their beliefs about (the parameters of) these acceptance functions (Fig. 2*B*). Beliefs may be inaccurate and lead proposers to offer too much or too little. However, through a process of observing offer acceptance and rejection by responders from different groups, uninformed agents can learn the group-specific acceptance thresholds held by responders from these groups (also see ref. 15).

To model this learning process formally, we adapted a variational approximation of the optimal Bayesian preference learner (14) (*Materials and Methods*). Beliefs about the parameters of the acceptance functions (θ_1, θ_2) were formalized as noisy Gaussian distribution with adjustable means (μ_1, μ_2) and SD (Σ_1, Σ_2). The variational Bayesian approach analytically derives how the belief distribution parameters are adjusted following observations of responder's decisions to accept or reject a given offer. Like (simpler) reinforcement-learning algorithms, the trial-by-trial updating of the mean value of the belief function parameters is governed by a choice prediction error – the difference between the observed choice and the expected probability of the offer being accepted.

Like the static case described above, proposers' beliefs about the acceptance function are used to form expected gain functions over the available offers, and which offer to make is governed by a softmax decision function. Simulations confirmed that such a model efficiently recovers group-specific acceptance function parameters (*Materials and Methods*; Fig. 2 *C* and *D*). Note that

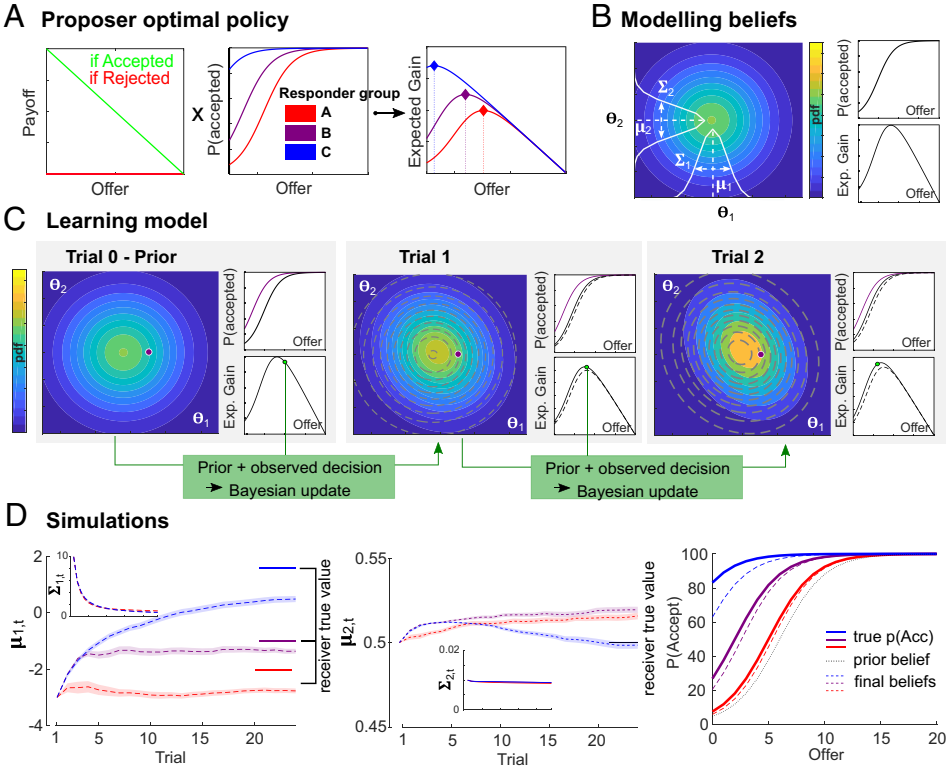

Fig. 2. Adapting offers to different acceptance thresholds. (*A*) Proposer optimal policy. The ultimatum payoff structure (*Left*) is combined with (known) acceptance functions (*Middle*) to derive expected gain from each possible offer and for each responder group (*Right*). The optimal policy is to make the offer with the maximum expected gain (diamonds). (*B*) Modeling beliefs about acceptance thresholds. Assuming proposers do not know the responders' true acceptance function parameters, but act on an internal representation – belief – that takes a Gaussian form $p(\theta) = N(\mu,\Sigma)$. The colored surface represents the belief multivariate probability density function with the marginal probability distribution for each parameter represented as white curves. (*C*) Modeling the learning of acceptance thresholds. Consider a proposer, represented by their belief probability density function [pdf; colored surface represents $p(\theta)$], confronted with a specific group, represented by the parameters of its acceptance function (purple dot). At each trial, the proposer uses their estimated acceptance function (black curve; *Top Right Insets*) to produce an expected gain function (*Bottom Right Insets*) and makes an offer that (soft)maximizes expected gain (green dot). In this case, the offer is accepted, and the proposer uses this information to update their beliefs using Bayes' rule. The peak of the belief probability density function (pdf) gets closer to the true parameters (purple dots). Dotted lines represent the previous trial features. (*D*) Simulating agents employing the learning model. Simulations (N = 100) show that the Bayesian model converges to the intercept (*Left*) and slope (*Middle*) of different groups' acceptance functions (color codes are identical to panel *A*). The right panel shows the original (dotted black line) and final estimated (dotted colored lines) acceptance functions, with true acceptance function superimposed (thick colored lines). For all simulations, we used $\beta = 2$; $\mu_{1,0} = -3$; $\mu_{2,0} = 0.5$; $\Sigma_{1,0} = 10$; $\Sigma_{2,0} = 0.01$.

responder groups mainly differ in the average offer that they accept (i.e., the belief function's mean intercept $\theta_{1,t}$). The learning process therefore mainly updates the belief function's mean intercept ($\mu_{1,t}$), and reduces the uncertainty around the belief function parameters ($\Sigma_{1,t}$, $\Sigma_{2,t}$). Over trials, Bayesian learning allows initially uninformed proposers to form accurate beliefs about the acceptance function parameters held by responders from different groups, and to select offers that (soft)maximize expected payoff.

Consistent with the standard economic theory, we thus far assumed the (simulated) proposers' only objective is to maximize expected personal gain. Yet, extensive literature documents that, in addition to concerns about personal gain, people are concerned also with others' outcomes (2, 9, 10, 16, 17). For example, people often prefer equal rather than unequal wealth distributions, even when this is personally costly (7, 9, 10). Adding such inequality aversion to the utility function that governs the decisions of our Bayesian model reveals how the optimal proposer policy changes (Fig. 3*A*; *Materials and Methods*). Specifically, fully informed agents with inequality aversion should make higher offers than strict gain-maximizing agents. Uninformed agents with inequality aversion can also leverage the Bayesian update rule to adapt their offers to the unknown acceptance thresholds of the different groups. Although the changes made to account for inequality aversion are limited to the decision part of the model and the belief-updating algorithm remains computationally the same, simulations show that adding inequality aversion biases the learning of the acceptance function intercept ($\mu_{1,t}$; Fig. 3*B*). Although belief updating follows optimal Bayesian learning, inequality aversion prevents agents to sample informative responders' actions (*SI Appendix*, Section II.5 for a quantification of this intuition). Over repeated encounters with new partners from different groups, this leads to a misrepresentation of the acceptance function parameters for groups with low acceptance thresholds ($\mu_{1,t}$; Fig. 3*B*) — agents with social preferences end up believing that responders from some groups require more generous offers than is necessary (and more generous than they would believe when repeatedly bargaining without inequality aversion; $\mu_{1,end}$; Fig. 3 *B* and *C*).

The behavioral consequences of this learning bias are twofold: First, offers made by proposers are consistently higher in the presence of inequality aversion, and higher than needed to reach agreement especially when interacting with responders from groups with low acceptance thresholds; Second, the subjectively estimated probability of an offer being accepted at the end of learning is lower in the presence of inequality aversion, especially for responders from groups with low acceptance thresholds and for low offers (Fig. 3*C*). In short, Bayesian belief updating in conjunction with inequality aversion predicts the emergence of misperceptions of

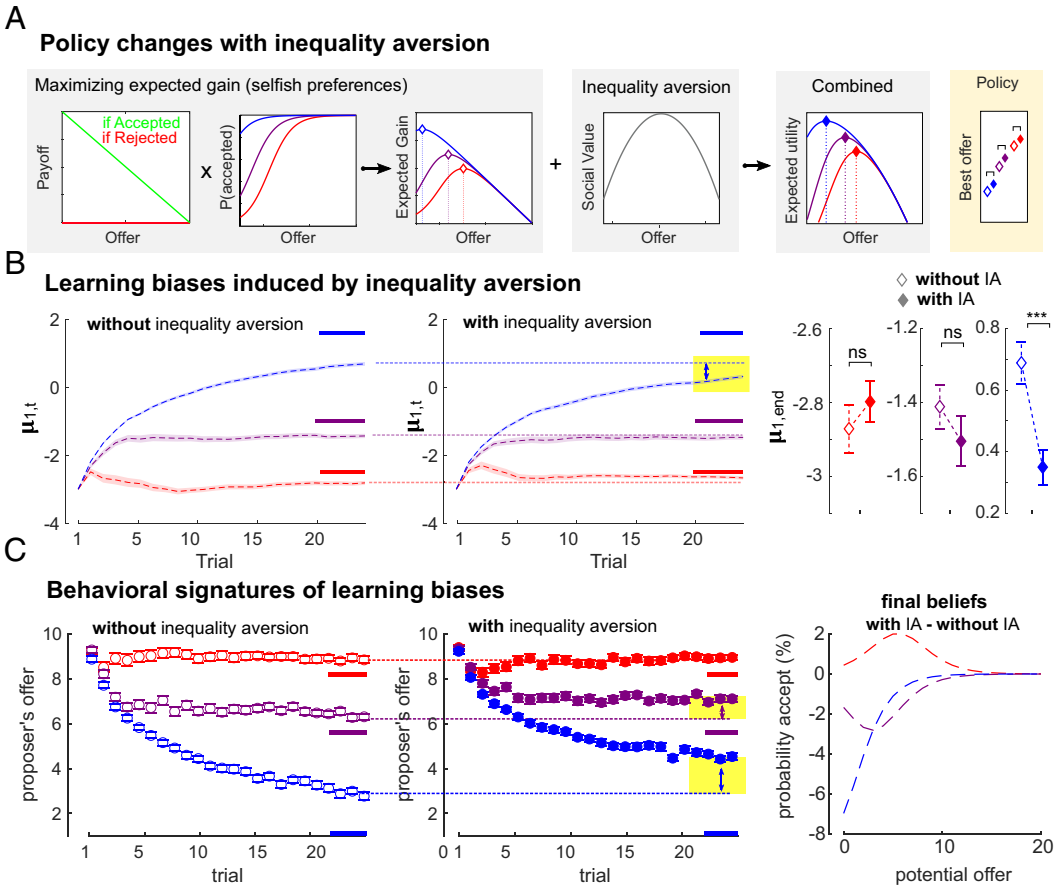

Fig. 3. Inequality aversion biases the learning of unknown acceptance thresholds. (*A*) Optimal policy for proposers with and without inequality aversion. The ultimatum payoff structure is combined with acceptance functions to derive expected gain for each possible offer and for each group (*Left*). Adding an inequality aversion (IA) term (*Middle*) generates an expected utility function that includes inequality aversion (*Right*), changing what offer has maximum expected utility (diamonds). (*B*) Simulating learning in inequality averse agents. Two sets of simulations (N = 100 each) were performed with (*Middle*) or without (*Left*) the inclusion of inequality aversion (IA). After 24 trials, Bayesian learning converges to different beliefs about the responders' acceptance function intercepts (μ_1) when IA is included, especially for the most lenient group that would accept relatively low offers (blue color – *Right*). For all simulations, we used $\beta = 2$; $\mu_{1,0} = -3$; $\mu_{2,0} = 0.5$; $\Sigma_{1,0} = 10$; $\Sigma_{2,0} = 0.01$; $\omega = 3$. (*C*) Behavioral consequence of inequality aversion. During learning, inequality aversion increases offers (*Left* vs. *Middle*). After learning, proposer's posterior beliefs differ depending on whether inequality aversion was present during learning, or not (*Right*). Shown is the difference between the estimated (posterior) probability of acceptance after learning with (versus without) inequality aversion, for the three responder groups. ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$, $^{\#}P < 0.10$ (two-tailed tests).

the partner's acceptance threshold, leading agents to make offers that i) are higher than necessary to secure agreement, and ii) forego diagnostic information about their responder's acceptance threshold.

Eliciting Proposers' Behavior. We tested model predictions in preregistered and fully incentivized experiments (total $N = 420$, *Materials and Methods*). Participants made offers to responders that were identified by three neutral symbols, akin to group-specific identity markers such as language or clothing (Fig. 4*A*). Unbeknownst to proposers, symbols corresponded to responder groups with a particular acceptance threshold (per Fig. 1*A*). Specifically, participants played multiple single-shot ultimatum games against different responders from the three different groups. For each of these interactions, participants could only identify the group that the responder belonged to. Accordingly, across interactions with responders from these different groups, participants could learn and adapt to their partners' group-specific acceptance thresholds.

To address how inequality aversion influences learning, we exposed participants to two blocks of trials. In one block of trials, participants faced human responders whose earnings depended on their offers (social condition). In another block, participants faced computer agents that participants knew were behaviorally identical to human responders (nonsocial condition; *Materials and Methods*) (11). Thus, whereas in the social condition, responder acceptance of ultimatum offers affect both the proposer and the responder, in the nonsocial condition, responder decisions only influenced the proposer's payoff. Because acceptance functions are identical, proposers' gain-maximizing strategy should be identical between social and nonsocial conditions, and inequality aversion should play a role in the social condition only (6, 10). Indeed, participants made higher initial offers and higher offers on average in the social compared to the nonsocial condition (collapsed across experiments: initial offer: $b \pm \text{SE} = 0.732 \pm 0.139$, $P < 0.001$; average offer: $b \pm \text{SE} = 0.296 \pm 0.032$, $P < 0.001$). Both here and for the other behavioral results reported below, we find evidence for this effect when collapsing across experiments, and for each experiment independently (*SI Appendix*, Table S1 and Fig. S3 for details). Furthermore, we could rule out that risk-preferences were involved in initial offers in the social condition more than in the nonsocial condition (17) (*SI Appendix*, *Section I.7*).

As anticipated in our model simulations, participants progressively learned, across repeated offers, their responders' group-specific acceptance thresholds (i.e., responder group final offer: $b \pm \text{SE} = -1.343 \pm 0.047$, $P < 0.001$; Fig. 4 *B* and *C*). Importantly, however, participants' final offers to the different responder groups differed depending on whether they interacted with human versus computer responders (nonsocial/social × responder group: $b \pm \text{SE} = 0.103 \pm 0.047$, $P = 0.030$; Fig. 4 *B* and *C*). Again, results replicated in each individual experiment (*SI Appendix*, Section I.6).

Validating the Model: Learning Behavior. The data from Experiments 1 to 5 allowed us to quantitatively validate our computational theory, through model comparison and falsification. To jointly demonstrate that the learning behavior is well accounted

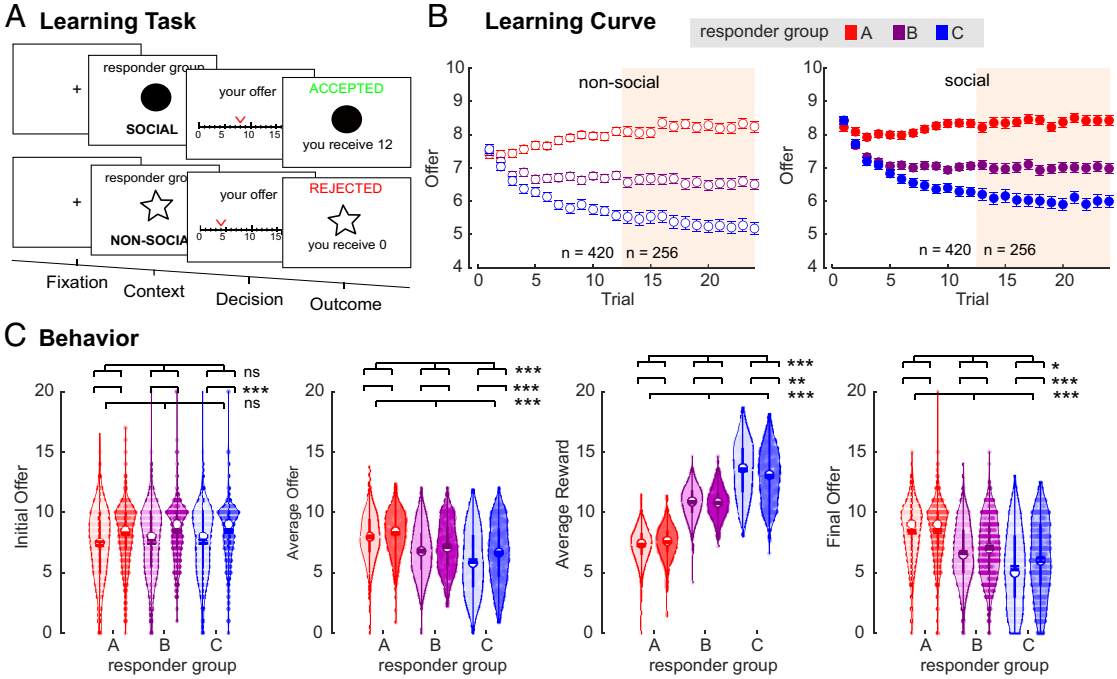

Fig. 4. Experimental participants learning group-specific acceptance thresholds. (*A*) Trial timeline of the behavioral experiments. In alternating blocks, either with human responders or behaviorally identical computer agents, participants repeatedly made ultimatum offers to responders sampled from three different responder groups, with group membership indicated by a neutral shape that allowed participants to track group-specific accept/reject histories. Responders' acceptance functions were unknown to participants and decisions only had social consequences when interacting with human (versus computer) responders. (*B*) Offers over trials. Across multiple encounters, participants' offers converge on the acceptance thresholds of the three different responder groups, depicted with different colors (red = responder group A; purple = responder group B; blue = responder group C), and in the two different conditions (*Left*: nonsocial; *Right*: social). Dots represent mean ± SE. Convergence of offers on responder acceptance thresholds is impeded when interacting with human rather than computer-simulated responders (shown mean ± SE). The light orange shaded area is based on 256 subjects (Experiments 2, 3, and 5 that completed 24 trials per responder group and block); white area indicates data points with the full sample (Experiments 2 to 5; $N = 420$; covering 12 trials per responder group and block). (*C*) Offers and outcomes. Violin plots depict the sample behavior for initial offer (*leftmost*), average offer (*Middle-left*), average earnings (*Middle-right*), and final offer (*rightmost*). The light versus dark colored violins plots and dots indicate that participants were interacting with computer (nonsocial condition) versus human responders (social condition), for the three different responder groups. White dots within the violin plots represent the sample median, the error bar indicates the mean ± SE and the thick central line indicate the 25 to 75% quantiles. Light-colored dots represent all individual datapoints. ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$, $^{\#}P < 0.10$ (two-tailed tests).

for by our Bayesian updating model, and that the observed effect of the social condition is satisfactorily captured by the weight of inequality aversion on individual decisions, we compared two models. Both shared the Bayesian preference learning algorithm describe previously. Yet, whereas the null model had a utility function that only integrated monetary gain maximization and omitted inequality aversion, the alternative model assumed a utility function that additionally incorporated inequality aversion in the social condition only, where not only proposers but also responders are affected by bargaining outcomes.

We first checked that the parameters of the model with inequality aversion were estimable, and that both models were identifiable in a model comparison analysis (*SI Appendix*, *Section II.1-3*). Fitting both models to our participants' data using Bayesian Model Selection analysis, identified the model with inequality aversion as the model that best accounted for the patterns of offers observed in the data (18) (Protected Exceedance Probability = 100%; Fig. 5*A*).

We then extracted the inequality aversion model's posterior predictive fits (i.e., trial-by-trial estimate of expected offer), and found that they closely match the trajectory of participants' offers in all conditions (Fig. 5 *A* and *B*). At the individual level, those posterior predictive fits accounted for a variety of different strategies, which further validated the ability of this model to flexibility account for most patterns of behavior observed in the task (*SI Appendix*, *Section II.4-5*).

Because model comparisons are necessary but not sufficient to validate a model (19), we also performed model falsification by simulating synthetic data using both models and the parameters estimated from the participants' data (*Materials and Methods*). When comparing behavioral patterns obtained from the synthetic data to those observed in the participants, we observed that the model with inequality aversion closely mirrored all key behavioral patterns, whereas the null model did not (Fig. 5*C*). This suggests that our model parsimoniously explains our participants' (biased) learning of

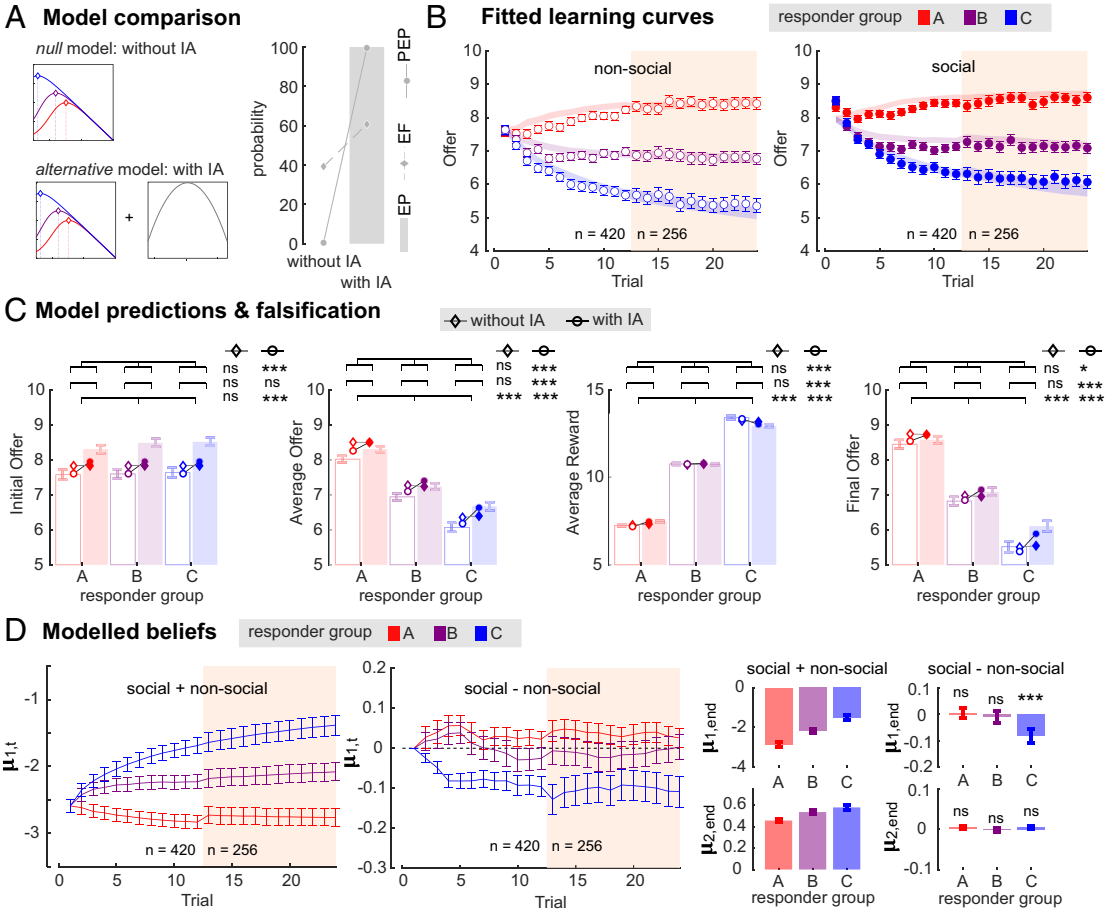

Fig. 5. Modeling the convergence of participants' offers on responder acceptance thresholds. (*A*) Model comparison. Using Bayesian model comparison, we contrasted a null model (Bayesian learning without inequality aversion) and the alternative model that includes inequality aversion (IA) in the utility function for the social condition. The model with inequality aversion was superior in terms of Exceedance Probability (how likely a model within the testing set is more frequent than any other; EP, histogram), Protected Exceedance Probability (PEP, diamond), and Expected Frequency (EF, dots). (*B*) Posterior predictive fits. The shaded areas represent the model fits (m ± SE) overlayed over participants' offers (dots and error bars: m ± SE). The model with inequality aversion reproduced the convergence on acceptance thresholds observed in the nonsocial (*Left*) and social (*Right*) conditions for the three different responder groups (red = responder group A; purple = responder group B; blue = responder group C). The light orange shaded (white) area is based on 256 subjects from Experiments 2, 3, and 5 (full sample with $N = 420$). (*C*) Model falsification. We simulated offers using both the null model (diamonds) and the model with inequality aversion (dots), using the parameters estimated from the participants' data. The model with inequality aversion reproduces the patterns observed in the participants' behavior (light-colored bars) for initial offer (*leftmost*), average offer (*Middle-left*), average reward (i.e., earnings; *Middle-right*) and final offer (*rightmost*), in the nonsocial (open markers) and social (filled markers) conditions, and for the three different responder groups. (*D*) Modeled participant beliefs. The time-course depicts the change in beliefs about the acceptance function's intercept ($\mu_{1,t}$), derived from fitting participants' offers to the model with inequality aversion. Shown are model fits collapsed across nonsocial and social conditions (*Left*) and as a function of the difference between the social and nonsocial conditions (*Right*), for the three different responder groups (red = responder group A; purple = responder group B; blue = responder group C). The light orange shaded (white) area is based on 256 subjects from Experiments 2, 3, and 5 (full sample with $N = 420$). The bars show the belief about the acceptance function's parameters at the end of learning (intercept $\mu_{1,end}$, *Top*; and slope $\mu_{2,end}$, *Bottom*) in all conditions (both nonsocial and social; *Left*) and as a function of the social condition (social – nonsocial; *Right*). Fitting model simulations, results show a specific difference between social versus nonsocial in the belief intercept for the most lenient responder group (blue = responder group C). ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$, $^{\#}P < 0.10$ (two-tailed tests).

group-specific rules of engagement (Fig. 5 *B* and *C* and *SI Appendix*, *Section II.3*). Finally, we extracted the latent variable from the model with inequality aversion fitted to the participants data, i.e., the underlying time series of their beliefs over the acceptance function intercept ($\mu_{1,t}$). As expected, participants' modeled beliefs about the different responder groups' acceptance threshold converged toward the true value, regardless of responder type (social + nonsocial; Fig. 5*D*). However, when contrasting conditions, we find that participants gradually overestimated this threshold when facing human responders from the group with the most lenient acceptance threshold (i.e., $\mu_{1,t}$ becoming more negative; social – nonsocial; Fig. 5*D*, *Middle*). Accordingly, the modeled beliefs that best account for our participants' behavior show the bias predicted by our Bayesian model with inequality aversion — inequality aversion induces inefficient sampling of responders' actions and participants fail to accurately learn another group's acceptance thresholds (when interacting with human rather than computer responders).

Validating the Model: Posterior Beliefs. Although the modeled beliefs of our participants confirm our hypothesis, a more direct test is to measure posttask beliefs directly from participants. To this end, participants in several of our preregistered experiments (N = 364) performed an incentivized belief estimation task to elicit their beliefs about their responders' acceptance functions (Fig. 6*A*; *Materials and Methods*). Results showed that participants' beliefs about the acceptance probabilities of the different responder groups were as predicted, with estimated acceptance function intercepts that monotonically increased with the group's true acceptance threshold (social + nonsocial, Fig. 6 *B* and *C*). Importantly, participants estimated the acceptance thresholds of human responders to be higher than those of behaviorally identical computer agents, specifically for the most lenient responder group C ($t(363) = -2.290$, $P = 0.023$, Fig. 6 *B* and *C* and *SI Appendix*, Section I.8). Accordingly, when interacting with responders from groups with lenient acceptance thresholds, inequality aversion leads participants to over-estimate what responders from these groups would accept. In fact, extracting the difference between social and nonsocial conditions in participants' beliefs about the acceptance probabilities over all offers shows that participants end up believing that responders require more generous offers than necessary, and this is particularly true for responders from the lenient group, and for lower offers (Fig. 6*D*).

Neuroimaging: Validating the Model. Thus far, we showed that modeling humans as Bayesian learners with inequality aversion accurately accounts for behavior and belief-updating both during

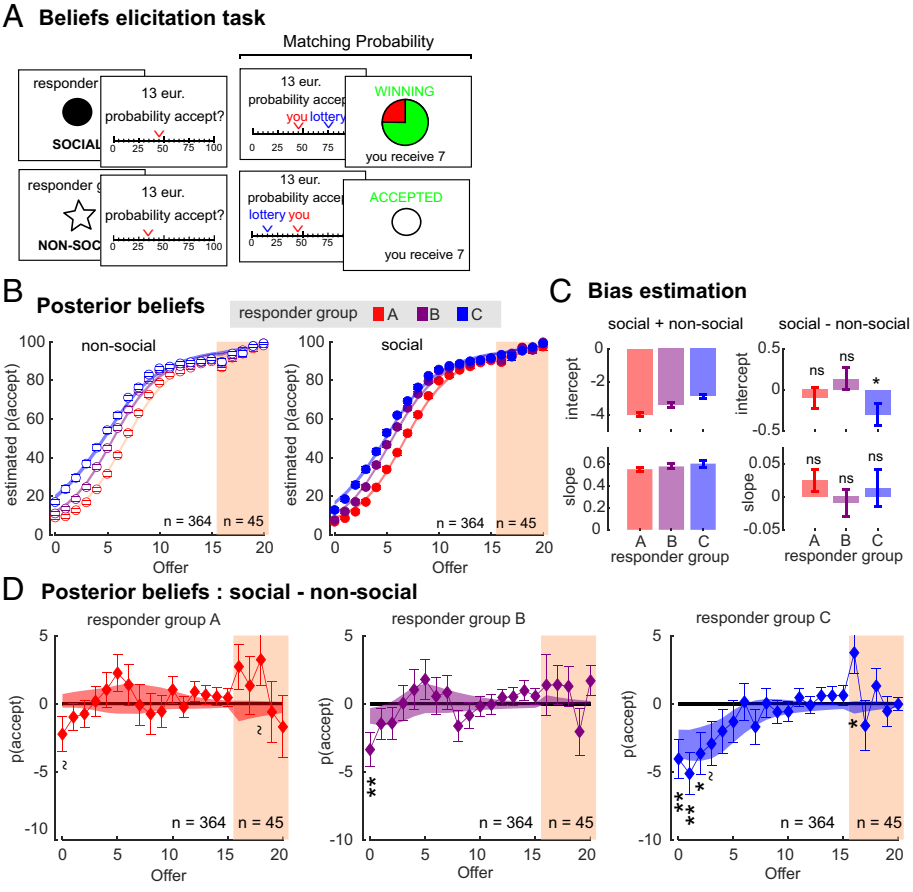

Fig. 6. Making offers that have social consequences bias beliefs. (*A*) Belief elicitation task. After making offers, and for each possible offer, proposers indicated the subjective probability of acceptance by human and (behaviorally identical) computer agents from each of the three responder groups. (*B*) Posterior beliefs. Proposer answers were averaged (dots with error-bars) and fitted to sigmoid functions (lines) that represent proposers' estimates of each responder group's acceptance function (red = responder group A; purple = responder group B; blue = responder group C). Beliefs about human (computer) responders are shown left (*Right*). The light orange shaded area is based on 49 participants (Experiment 5) and the white area is based on the full sample (N = 364). (*C*) Bias estimation. Intercepts and slopes of the sigmoid fits to the participants' beliefs were tested for differences between human and computer agents in each of the three responder groups. The bars show the values of the function parameters for the elicited beliefs (intercept, *Top*; and slope, *Bottom*) in all conditions (both nonsocial and social; *Left*) and as a function of the social condition (social – nonsocial; *Right*). Fitting model simulations, results show a specific difference between social vs. nonsocial in the belief intercept for the most lenient responder group (blue = responder group C). (*D*) Effects of interacting with human participants on posterior beliefs. The graphs depict the difference between the estimated probability of acceptance by human and (behaviorally identical) computer agents from each of the three responder groups in participants (diamonds and error bars) and corresponding sigmoid fits (shaded areas). ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$, $^{\#}P < 0.10$ (two-tailed tests).

and after learning. Underlying our model is the assumption that: i) participants' updating processes can be captured by an approximation of (ideal) Bayesian learning that leverages a choice-prediction error; ii) the biasing effect of the social condition operates when making an offer (rather than when processing offer acceptance/rejection); and iii) participants make decisions by maximizing a utility function that linearly combines monetary gain expectations and, in the social condition, inequality aversion. We examined these assumptions using functional neuroimaging ($N = 49$; *Materials and Methods* and *SI Appendix*, Section III) and modeled BOLD activity separately for the decision and outcome steps, and for the social and nonsocial conditions. We additionally included parametric modulators for variations in expected utility during decision-making, and for choice prediction error during feedback (*SI Appendix*, Section III.4.

With regard to the proposed learning mechanism, at the whole-brain level, choice prediction errors positively correlated with BOLD in the ventral striatum (VS), the ventromedial prefrontal cortex (VMPFC), the orbitofrontal cortex (OFC), and precuneus. Choice prediction errors negatively correlated with BOLD in the dorsal anterior cingulate (dACC) and anterior insula (Fig. 7*A*). These two networks have been repeatedly associated with reward- and punishment-driven reinforcement learning (20), and suggest that choice prediction error tracks the actual underlying updating process used by our participants. This was further confirmed when we computed a general linear model that related the choice prediction, and the responder's actual decision, to BOLD in the VS (21) (*SI Appendix*, Section III.4). VS activity correlated negatively with choice prediction [$t(48) = -2.354$, $P = 0.0227$], and positively with the responder's decision [$t(48) = 6.894$, $P < 0.001$; Fig. 7*B*]. We note that this finding was specific to choice prediction and not true for the related reward prediction errors, that were marginally positively correlated with VS BOLD ($t(48) = 1.829$, $P = 0.0736$; Fig 7*B*) *SI Appendix*, Section III.4.

Whereas our model's learning signal—choice prediction—robustly tracked neural activity in a well-validated learning region (VS), effects of our social vs. nonsocial manipulation should manifest mostly in the utility function, and thus correlate with neural activity when participants decide what to offer (rather than in the feedback phase when participants update their beliefs). Whereas simply contrasting the social and nonsocial conditions did not show significant differences in BOLD, a more sensitive multivoxel pattern analysis showed significant differences between conditions in the dorsal anterior cingulate cortex (dACC) and the superior temporal sulcus (STS), regions often associated with updating strategic choices (22), and with social cognition and theory of mind (23) (Fig. 7 *C* and *D*). As expected, this condition effect was stronger during decision-making than during feedback. Finally, when we split the expected utility of the considered offer into expected gain and (in the social condition) inequality aversion, we find that neural activity in the ROIs previously identified as correlating with expected utility (VMPFC, VS, Insula and dACC) encoded expected gain in both social and nonsocial conditions (positively for VS and VMPFC and negatively for Insula and dACC; all ps < 0.05) (Fig. 7 *E* and *F*), but that only the Insula and dACC additionally encoded inequality aversion (in the social condition; *SI Apendix*, Section III.4–5; also for replication at the whole-brain level). This functional dissociation at the neuroanatomical level provides additional evidence for the idea that expected gain and inequality aversion are separable components of the expected utility function.

Discussion

Through repeated interactions with people from different groups, individuals learned their partners' implicit rules of engagement, and updated their beliefs accordingly. Presumably because of inequality aversion, individuals made higher offers when interacting with human rather than computer agents and did not fully explore the consequences of making more self-serving offers to human responders. This restrictive sampling of their available option space not only resulted in higher offers than were required to reach agreement, but also in inaccurate beliefs about "what it takes to agree".

Whereas social preferences and beliefs are often conceptualized as distinct components in decision-making models, our results show, first, that social preferences and beliefs are not independent and, second, how they can coevolve. This not only contradicts a central assumption of standard economic theory that preferences and beliefs are independent, but also provides a previously unidentified mechanism underlying the development and persistence of group-based perceptions and (false) stereotypes. Specifically, inequality aversion influenced what offers individuals made, and which ones they omitted. With responders acting on their group-based acceptance thresholds, some offers were more likely to be accepted than others; yet some offers were rarely made, and proposers could thus not learn whether or not responders from certain groups would have accepted these. The result was not only that offers were often more generous than needed, but also that proposers developed predictably wrong beliefs about what is typically acceptable in other groups. False stereotypes develop partly because individuals' social preferences bias their behaviors toward unknown outsiders.

Our analysis and experiments revealed that sometimes people are more generous than strictly needed, allowing them to cooperate and reach agreements with individuals from distinctly different groups. This result counters the idea that intergroup relations are bound to gravitate toward negative misperceptions and hostility (24, 25). In contrast, because inequality averse individuals shy away from making self-serving offers, they may promote rather than hinder cooperation with groups that hold different norms and rules of engagement. For example, being inadvertently more generous with foreigners than one is expected to be (e.g., when tipping) can establish common ground for future cooperation. It may also lead to the erroneous belief that such generosity is expected. Crucially, distorted beliefs and expectations about "what it takes" avoids conflict and may foster rather than hinder cooperative exchange across group boundaries.

Implications for interactions across group boundaries and cultural divides should take into account, first, that estimating (generalizable) individual traits from behavioral tasks and fitted models is difficult as individuals do not consistently follow the same decision strategy across different tasks (see e.g., (26, 27) for the case of risk attitudes). Second, our participants originated from Western Europe and the United States. Inequality aversion may not be a universal human trait (5), and effects found may be specific to a particular sample. At the same time, the core insight that social preferences shape, over repeated interactions, how individuals learn the rules of engagement endorsed by people from other cultures may not be limited to a specific cultural setting. Indeed, our modeling of humans as Bayesian learners provides a flexible framework for understanding cross-cultural learning and belief updating, and to predict what inhibits cross-group cooperation and agreement. Research can be designed to invoke and test different fairness considerations held,

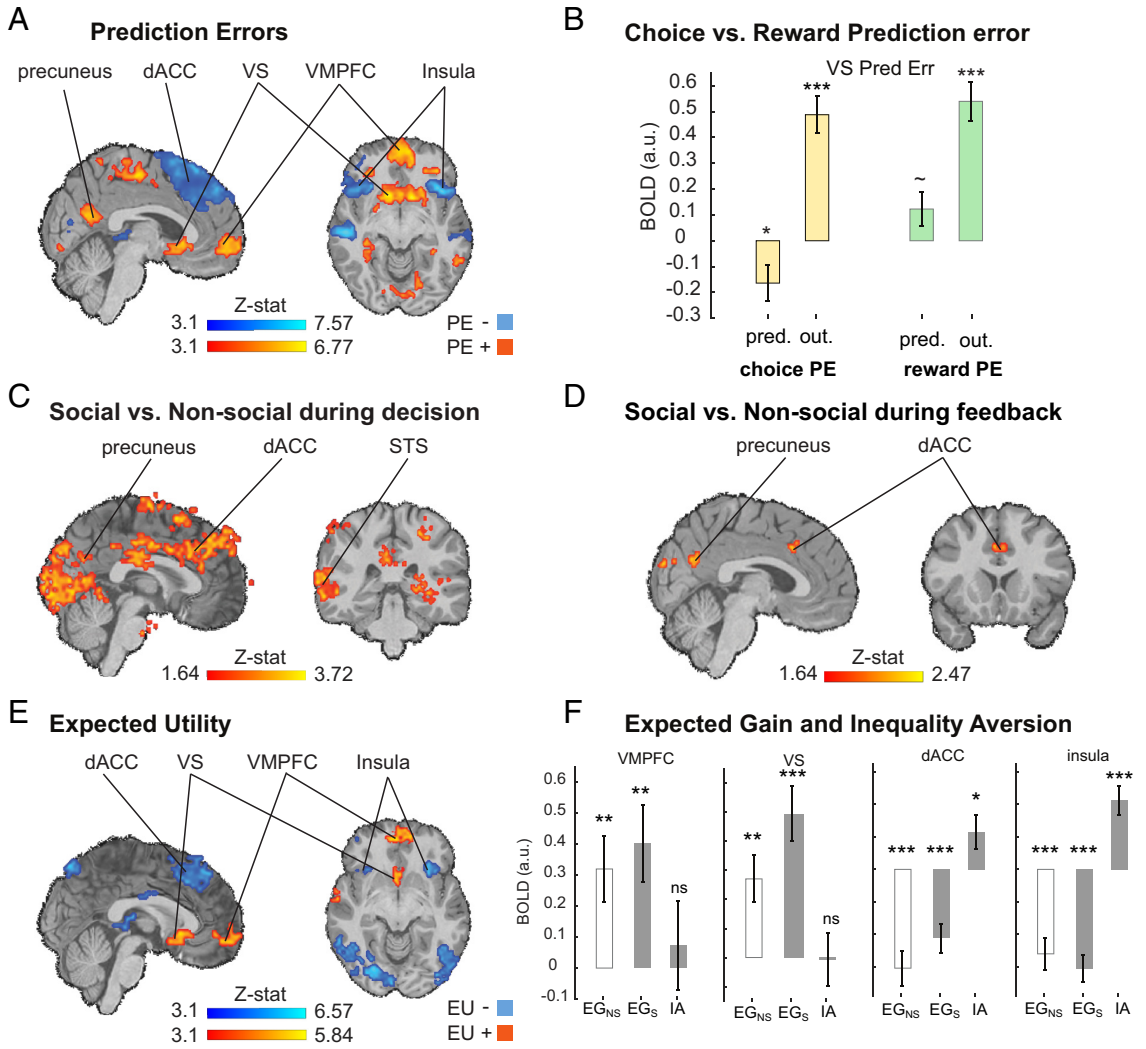

Fig. 7. Neuroimaging results. (*A*) Neural correlates of choice prediction error. The activation maps picture whole-brain significant BOLD for positive (orange/red) and negative (blue) correlations with choice prediction error, overlaid on an anatomical T1 (standard cluster forming threshold $Z > 3.1$ and cluster significance at $P < 0.01$). (*B*) Choice vs. reward prediction error in the ventral striatum. Bars depict the sample m±se of parameter estimates extracted from anatomical ventral striatum ROI for GLM prediction errors split into prediction (two left bars) and outcome (two right bars). Shown are our model's choice prediction error (white)—the signal our model uses to update predictions—and a reward prediction error (gray) that our model does not use to update predictions. Ventral Striatum BOLD correlates negatively with choice prediction and positively with accept/reject outcome. This was not true for reward prediction error. (*C* and *D*) Social concerns in the brain. Multivariate contrasts revealed more widespread differences during decision-making (*C*) than during feedback (*D*). (*E*) Neural correlates of expected utility. Activation maps picture whole-brain significant activations for positive (orange/red) and negative (blue) correlations with expected utility, overlapped on an anatomical T1. Bars depict the sample m ± SE of regression coefficients extracted in our ROIs (VMPFC, VS, Insula, dACC) for expected gain in the social (EG_S; colored-filled) and in the nonsocial (EG_{NS}; white-filled) conditions, and for inequality aversion in the social condition (IA; colored-filled). $^{***}P < 0.001$, $^{**}P < 0.01$, $^{*}P < 0.05$, $^{\#}P < 0.10$ (two-tailed tests).

for example, in non-Westerns contexts. In addition, future studies can incorporate other "moral sentiments" than inequality aversion, like empathy and guilt aversion. While these moral sentiments align when people make more or less selfish ultimatum offers and may thus be nonseparable when learning "what it takes", this can change when individuals decide whether to "return the favor" (i.e., as trustees in the standard Trust Game; (28)). In such situations, individuals may base their decision on a desire for equality, or to avoid feelings of guilt that emerge when violating what is normatively expected. There is evidence that in such situations of trust and reciprocity, guilt aversion recruits different neural circuitry than inequality aversion (28). Our model of Bayesian learners with social preferences can be useful to examine the enticing possibility that guilt aversion drives learning 'how to return the favor' distinctly different than inequality aversion.

Our neuroimaging results showed that activations of the negative (or "salience") network (Insula, dACC) were more robust than those of the positive network. Possibly, participants generally expect to make successful offers and the most salient and informative events are offer rejections, and this could explain the comparatively strong activation in the Insula and dACC (20). Relatedly, we did not anticipate that BOLD activity in the Insula and dACC correlated with the two components of the expected utility function (expected gain and inequality aversion), and that activity in dACC additionally discriminated the social vs. nonsocial conditions. In ultimatum bargaining, however, the Insula is involved in empathy and fairness (29, 30), and the dACC relates to social inferences, norms, and expectation violations (31, 32). As such, results corroborate that cooperation recruits neurocomputational processes in interrelated brain networks involved in value-based learning and decision-making, and in social cognition (32, 33). And, as shown here, these neurocomputational processes are shaped by social preferences and can give rise to inaccurate beliefs about what it takes. Ironically, perhaps,

such inaccurate beliefs about what others expect may enable humans to interact with culturally distinct others in a constructive and mutually beneficial way.

Materials and Methods

Computational Modeling and Simulations. We modeled a "Bayesian Preference Learner" (BPL) (14). As BPL, proposers assume that their responder's choices obey a softmax decision rule. This transforms a linear utility function of offers into a probability of accepting an offer (Fig. 1*B*). This decision rule includes a potential offer: $O \in [0{:}\,20]$, a sigmoid function representing the responder's acceptance function $s{:}\ x \rightarrow 1 / \left(1 + \exp(-x)\right)$ with intercept θ_1 and slope θ_2, the responder's binary choice at trial t: $a_t \in \{0, 1\}$, and the estimated (linear) utility that responders derive from an offer O: $f(\theta, O) = \theta_1 + \theta_2 O$. Here, learning about other's attitude toward fairness reduces to updating one's belief about the parameters θ of the acceptance function. Because it fully characterizes responders' behavior, we refer to θ as responders' *trait*. The estimated probability that the responder accepts an offer O can then be written as:

$$p\left(a_t = 1 \mid \theta, O\right) \triangleq s(f(\theta, O)). \qquad [\mathbf{1}]$$

Before having observed any responder decision, proposers are endowed with some prior belief $p(\theta)$ about the responder group trait θ. Without loss of generality, we assume that this prior belief $p(\theta) = N(\mu_0, \Sigma_0)$ is Gaussian with mean μ_0 (which captures the direction of the proposer's bias) and variance/covariance Σ_0 (which measures how uncertain the proposer's prior belief is; Fig. 1*D*). Observing the responder's accept or reject decision gives the proposer information about θ, which can be updated trial after trial using Bayes' optimal probabilistic scheme: $p\left(\theta \mid a_{\rightarrow t}\right) \propto p\left(a_{\rightarrow t} \mid \theta\right) p(\theta)$, where $p\left(\theta \mid a_{\rightarrow t}\right)$ is the proposer's posterior belief about the responder's acceptance function after trial t. To highlight the trial-by-trial, sequential (online) form of Bayesian learning, this can be rewritten as

$$p\left(\theta \mid a_{\rightarrow t}\right) \propto p\left(a_t \mid \theta\right) p\left(\theta \mid a_{\rightarrow t-1}\right). \qquad [\mathbf{2}]$$

In other words, after observing responder's decision a_t, the proposer can update her (posterior) belief about the responder's behavioral trait $p\left(\theta \mid a_{\rightarrow t}\right)$ by combining the likelihood of observing the decision $p\left(a_t \mid \theta\right)$ with her preceding belief about the responder's behavioral trait $p\left(\theta \mid a_{\rightarrow t-1}\right)$.

Eq. **2** can be approximated using a variational-Laplace scheme, which essentially replaces the integration implicit in Eq. **2** with an optimization of the sufficient statistics of the approximate posterior distributions (18). This gives semianalytical expressions for the trial-by-trial update rules of two first moments of the posterior probability density function. In brief, we approximate the posterior belief $p\left(\theta \mid a_{\rightarrow t}\right) \approx N(\mu_t, \Sigma_t)$ in terms of a Gaussian distribution with mean μ_t and variance Σ_t. Given the observed decision a_t to the offer O_t made at trial t, this leads to the following learning (update) rules for the belief about the responder's expectations/acceptance function:

$$\Sigma_t = \left(\Sigma_{t-1}^{-1} + s\left(f\left(\mu_{t-1}, O_t\right)\right)\left(1 - s\left(f\left(\mu_{t-1}, O_t\right)\right)\right) \nabla f|_{\mu_{t-1}}{}^T \nabla f|_{\mu_{t-1}}\right)^{-1},$$
$$\mu_t = \mu_{t-1} + \Sigma_t \nabla f|_{\mu_{t-1}} \left(a_t - s\left(f\left(\mu_{t-1}, O_t\right)\right)\right) \qquad [\mathbf{3}]$$

where $(o)\ \nabla f = \partial f / \partial \theta$ is the gradient of the linear utility function f (cf. Eq. **1**) with regard to the acceptance function's parameters θ. Critically, and paralleling simpler models in reinforcement learning, it can be seen from Eq. **3** that the change in the agent's posterior mean $\mu_t - \mu_{t-1}$ is driven by a choice prediction error $a_t - s(f(\mu_{t-1}, O_t))$, whose impact is modulated by the agent's subjective uncertainty Σ_t. Also, note that the proposer's posterior uncertainty about the responder's behavioral trait Σ_t is monotonically decreasing over trials. Iterated through time or trials, Eq. **3** essentially describes how the proposer learns about the responder's probability to accept any offer (Fig. 1*E*; further mathematical details can be found in ref. 14).

The free parameters of the learning module that can be adjusted/fitted to account for our participants' behavior are the prior beliefs about responder's intercept θ_1 and slope θ_2 of their acceptance function, with $\mu_0 = [\mu_{1,0}, \mu_{2,0}]$, where $\mu_{1,0}$ is the prior mean of the responder's intercept and $\mu_{2,0}$ is the prior mean of responder's slope; $\Sigma_0 = \begin{bmatrix} \Sigma_{1,0} & \Sigma_{12,0} \\ \Sigma_{12,0} & \Sigma_{2,0} \end{bmatrix}$, where $\Sigma_{1,0}$ is the prior variance of the responder's intercept and $\Sigma_{2,0}$ is the prior variance of the responder's slope. $\Sigma_{12,0}$ is the prior covariance between the intercept and the slope, which we assume is 0. Hence: $\Sigma_0 = \begin{bmatrix} \Sigma_{1,0} & 0 \\ 0 & \Sigma_{2,0} \end{bmatrix}$; however, Bayesian learning induces a nonzero posterior covariance between the intercept and slope (i.e., $\Sigma_{12,t} \neq 0$ for $t > 0$).

Given the responder's choices up to trial t, the proposer can now form a prediction about the other's probability to accept any offer O at trial $t+1$: $E\left[a_{t+1} \mid O_{t+1}, a_{\rightarrow t}\right] = E\left[s\left(f\left(\theta, O_{t+1}\right)\right) \mid a_{\rightarrow t}\right]$. For simplicity and for consistency with the modeling of posterior beliefs, we assume that predictions only depend on the parameter estimated mean μ_t.

$$E\left[a_{t+1} \mid O_{t+1}, a_{\rightarrow t}\right] = s\left(f\left(\mu_t, O_{t+1}\right)\right). \qquad [\mathbf{4}]$$

These predictions can be used to compute the expected payoff of a given offer, based on the ultimatum game payoff matrix:

$$EG\left[O_{t+1}, a_{\rightarrow t}\right] = \left(e - O_{t+1}\right) \times s\left(f\left(\mu_t, O_{t+1}\right)\right). \qquad [\mathbf{5}]$$

Performing this computation across the entire offer space allows responders to identify which offer O_{t+1} gives the highest expected gain (Fig. 1 *D* and *E*).

Inequality Aversion. We hypothesized that proposers do not simply try to (soft) maximize their expected payoff EG (per Eq. **5**), but rather a complex expected utility function EU that integrates an additional inequality aversion component. The expected utility of an offer (EU) is a weighted sum of the monetary utility term accounting for the expected gain/payoff of the offer (EG), and a social utility term that measures how much the offer mitigates inequality (IA). For computational parsimony and simplicity, we model this IA term for offer O as the inverse quadratic function centered on an equal split of the proposer's endowment e:

$$IA(O) = -(O - e/2)^2. \qquad [\mathbf{6}]$$

Given the payoff matrix of the ultimatum game and assuming inequality aversion, the probability that responders accept any offer O at trial $t+1$ can be computed as:

$$EU\left[O_{t+1}, a_{\rightarrow t}\right] = \left(e - O_{t+1}\right) \times s\left(f\left(\mu_t, O_{t+1}\right)\right) - \omega \times (O - e/2)^2, \qquad [\mathbf{7}]$$

where ω is a weighting parameter that captures the relative importance of both utility terms for each proposer. Therefore, a general utility function in our task can be written as:

$$EU\left[O_{t+1}, a_{\rightarrow t}\right] = \left(e - O_{t+1}\right) \times s\left(f\left(\mu_t, O_{t+1}\right)\right) - \omega \times I_S \times (O - e/2)^2, \qquad [\mathbf{8}]$$

where I_S is a condition indicator function which is set to 1 in the social condition in which proposers interact with human responders in our experiments (see below) and is set to 0 in the nonsocial condition in which proposers interact with (behaviorally identical) computer agents.

Decision-Making. As is customary in decision science, we assume that participants softmaximize expected utility, i.e., probabilistically select offers in proportion to their relative expected utility. This can be captured by the normalized exponential function that models the probability of selecting an offer x_i as:

$$p\left(O = x_i\right) = \frac{\exp\left(\beta \times EU\left(x_i\right)\right)}{\sum_{j=1}^{M} \exp\left(\beta \times EU\left(x_j\right)\right)}. \qquad [\mathbf{9}]$$

Here, β is the inverse temperature parameter.

Simulations. All model simulations (Figs. 1 E and F and 2 *B* and *C*) were performed with the following parameters: $\beta = 2$; $\mu_{1,0} = -3$; $\mu_{2,0} = 0.5$; $\Sigma_{1,0} =$

10; $\Sigma_{2,0} = 0.01$; $\omega = 3$. Fig. 1*E* features three trials of a single synthetic participant. Figs. 1*F* and 2*B* features 24 trials of 100 independent subjects. Differences between social and nonsocial conditions in Fig. 2*B* were assessed with paired *t* tests.

Model Fitting and Comparison. Models were fitted by finding the parameter values θM which minimize the negative logarithm of the posterior probability (nLPP). This term is computed as $nLPP = -\log(P(\theta M \mid D,M)) \propto -\log(P(D \mid M,\theta M)) - \log(P(\theta M \mid M))$, where $P(D \mid M,\theta M)$ is the likelihood of the data (i.e., the observed offer, D) given the considered model M and parameter values θM, and $P(\theta M \mid M)$ is the prior probability of the parameters. Practically, we used matlab's fmincon function, initialized at multiple, random starting points of the parameter space (10 iterations). We performed a Bayesian model comparison between a null model that did not include the inequality aversion term, and an alternative model that did include the inequality aversion term. For model comparison, we compute the Laplace approximation to model evidence (LAME) (34):

$$LAME = -\log(P(\theta_M D, M)) + \frac{n}{2}\log(2\pi) - \frac{1}{2}\log|H|, \qquad [10]$$

where n is the number of parameters in the model, and H is the Hessian of the nLPP (negative log-posterior probability) function. We used this measure for Bayesian model comparison (18) as implemented in the VBA toolbox (https://mbb-team.github.io/VBA-toolbox/), a procedure which estimates expected frequency and exceedance probability of each model. Expected frequency quantifies the probability that the data of any randomly selected subject were generated by a model. By comparing it to chance level, one obtains the exceedance probability, which measures the belief that a model is more likely than all other models. Following (35), we performed a model identification analysis to verify that the models were identifiable (*SI Appendix*, Section II.2).

Experiments

Creating Different Responder Groups. Upon arrival in the laboratory, participants (N = 210) were seated in individual computer-equipped cubicles. They provided written informed consent, and were fully debriefed afterward. They received €6.50 for participation and on average M = €2.42 from decision-making. The experiment did not involve deception and received ethics approval from Leiden University (*SI Appendix*, Section I.1).

The experiment started with a short explanation of the rules of the ultimatum game. Following two comprehension questions, participants indicated, for each possible offer made by a proposer with an endowment e = 20, whether they would accept or reject the offer (e.g., "you are offered 20 and the proposer keeps 0.", "you are offered 1 and the proposer keeps 19"; Fig. 1*A*). To induce different acceptance thresholds and underlying expectations of which offers are acceptable or not, we manipulated the starting endowment of responders (e = 0, e = 10 or e = 20) across blocks (Fig. 1*A*). Participants' accept/reject decisions for each offer and each group (i.e., starting endowment) were averaged to obtain the mean frequency with which each offer was accepted (*SI Appendix*, Fig. S2). We fitted sigmoid functions over the resulting distributions to obtain the acceptance threshold of the respective group of responders (Fig. 1*B*). These functions served as inputs for our computational modeling and simulations, and to provide feedback to our human proposers in the learning and updating experiments.

Learning and Updating Experiments. We performed four preregistered experiments (*SI Appendix*, Section I.1). One experiment was performed in our behavioral laboratory with N = 50, two were performed on-line through the Prolific platform, with N = 157 and 164 (total N = 420, 227 females), and one was performed in the fMRI scanner (N = 49, 35 females). The experiments did not involve deception and received ethics approval from Leiden University. Participants provided written informed consent, and were fully debriefed afterward. They received €6.50 (Experiment 2) €7.26/£6.25 (Experiment 3 to 4) and €20 (Experiment 5) for participation and on average M = €2.29 per block from decision-making.

All experiments were computer-guided and participants were tested individually. Procedures and materials were largely identical across experiments apart from slight differences in number of trials (*SI Appendix*, Section I.3 for details). Upon providing informed consent, participants read instructions for the ultimatum game and answered three comprehension questions. They then made ultimatum offers to human and computer responders (block design, with either two or four blocks of 36 to 72 trials per block, see *SI Appendix*, Section I.3). In the social condition, participants were instructed that they were interacting with responders who had received different starting endowments, but were not told what these endowments were. Hence, they were aware that different implicit fairness rules may apply across groups. In the nonsocial condition, participants were told that they were interacting with computer generated lotteries programmed to mimic the behavior of participants who had received different starting endowments, that is, with computers programmed to behave like humans. Indeed, the task and responder behavior were identical across conditions except that in the social condition, decisions not only affected the earnings of the proposer but also the responder, whereas in the nonsocial condition, decisions only affected the earnings of the proposer.

One trial from each block was selected at random for payment. In each block, participants interacted in an interleafed fashion with responders that made decisions according to their underlying acceptance function of their respective responder groups (per Exp. 1, and Fig. 1*C*). In Experiments 2 and 5, each responder group was marked with a neutral shape such as a circle or square, and all shapes were randomized for each participant and only used once such that each block consisted of completely novel shapes to avoid learning across blocks. In Experiments 3 and 4 we emphasized the human/computer contrast by denoting each responder group by a different colored gender-neutral human silhouette (social treatment) or a different colored slot machine (nonsocial treatment). All colors were randomized for each participant, and only used once such that each block consisted of completely novel colors.

Experiments 2 to 4 were self-paced. In each trial, participants faced one responder that was only identified by their group membership with a shape or color. They then had to decide how to split their endowment between them and this responder (their proposal; between 0 and 20). After this decision, they learned whether their proposal was accepted or rejected and how much units they earned for that trial. Hence, on each trial, participants played a single-shot ultimatum game, facing one subject that belonged to one of our three responder groups.

Experiment 5 was conducted in the fMRI scanner. Each trial started with a fixation cross (1.5 to 2.5 s), followed by a screen showing the shape denoting the responder group (2 to 3 s). Participants used a slider to select an offer between 0 and 20, after which they were shown a screen indicating whether or not the offer was selected and how many units they earned for that trial (2 to 3 s).

Measuring Posterior Beliefs. After one social and one nonsocial block (for Experiments 3 to 5), participants completed a fully incentivized belief estimation task (we incentivized accuracy with a matching probability/auction mechanism (36, 37) and selected one trial at random for payment). Participants estimated the acceptance probability of each offer for each responder group participants had been making offers to. On each trial, participants were presented with a shape corresponding to one of the responder groups from

the previous blocks as well as an offer between 0 and 20. They were asked to identify, on a scale from 0 to 100%, how likely the given offer was to be accepted by someone from that responder group. This allowed us to estimate postlearning beliefs about the underlying acceptance function of the different responder groups.

Neuroimaging. Neuroimaging (Experiment 5) was performed using a standard whole-head coil on a 3-T Philips Achieva MRI system at the Leiden University Medical Center. Participants completed four runs, during which 400 T2*-weighted whole-brain echoplanar images (EPIs) were collected (TR = 2.2 s; TE = 30 ms, flip angle = 80°, 38 transverse slices, 2.75 × 2.75 × 2.75 mm +10% interslice gap). The first five dummy scans were discarded to allow for equilibration of T1 saturation effects. After each functional run, a B0 field map was acquired. Additionally, a 3D T1-weighted scan was acquired (TR = 9.8 ms; TE = 4.6 ms, flip angle = 8°, 140 slices, 1.166 × 1.166 × 1.2 mm, FOV = 224.000 × 177.333 × 168.000).

Following preprocessing (*SI Appendix*, Section III.3), neuroimaging data were analyzed with FSL (Oxford Centre for Functional MRI of the Brain Software Library; www.fmrib.ox.ac.uk/fsl). For all general linear models (GLMs), at the first level (within participants within runs), each participants' blood oxygen level dependent (BOLD) data were spatially smoothed with 5-mm FWHM gaussian kernel, high-pass temporal filtered, film prewhitened, and convolved with the canonical double-gamma hemodynamic response function (*SI Appendix*, Section III.4—6).

Data, Materials, and Software Availability. All datasets and scripts generated and/or analyzed in this article are publicly accessible at Open Science Framework (https://osf.io/rkbev/) (38) and Openneuro (https://openneuro.org/datasets/ds004553) (39).

ACKNOWLEDGMENTS. This project was supported by funding from the European Research Council (ERC) (AdG agreement n° 785635) to C.K.W.D.D., a NWO VENI Grant (016.Veni.195.078) to J.G., and a SNSF Ambizione Grant (PZ00P3_174127) and an ERC Starting Grant (948671) to M.L.

Author affiliations: [a]Institute for Psychology, Social, Economic and Organisational Psychology, Leiden University, 2333 AK Leiden, the Netherlands; [b]Paris-Jourdan Sciences Economiques UMR8545, Economics of Human Behavior Group, Paris School of Economics, 75014 Paris, France; [c]Swiss Centre for Affective Sciences, Faculty of Psychology and Educational Sciences, Université de Genève, 1202 Geneva, Switzerland; [d]Sorbonne Université, Inserm Unite 1127, CNRS unite 7225, 75005 Paris, France; [e]Paris Brain Institute (ICM), Motivation Brain & Behavior (MBB) Lab, Pitié-Salpêtrière Hospital, 75013 Paris, France; [f]Institute of Psychology, Social and Economic Psychology, University of Zurich, 8001 Zurich, Switzerland; and [g]Amsterdam School of Economics, Center for Research in Experimental Economics and Political Decision Making, University of Amsterdam, 1018 WB Amsterdam, the Netherlands

1. E. Fehr, I. Schurtenberger, Normative foundations of human cooperation. *Nat. Hum. Behav.* **2**, 458–468 (2018).
2. E. van Dijk, C. K. W. De Dreu, Experimental games and social decision making. *Annu. Rev. Psychol.* **72**, 415–438 (2021).
3. E. Sapir, "The unconscious patterning of behavior in society" in *The Unconscious: A Symposium*, (Alfred A. Knopf, 1927), pp. 114–142.
4. M. J. Gelfand, Differences between tight and loose cultures: A 33-nation study. *Science* **332**, 1100–1104 (2011).
5. P. R. Blake *et al.*, The ontogeny of fairness in seven societies. *Nature* **528**, 258–261 (2015).
6. S. Debove, N. Baumard, J. B. André, Models of the evolution of fairness in the ultimatum game: A review and classification. *Evol. Hum. Behav.* (2016), 10.1016/j.evolhumbehav.2016.01.001.
7. J. Henrich *et al.*, "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behav. Brain Sci.* **28**, 795–815 (2005).
8. A. Romano, J. Gross, C. K. W. De Dreu, Conflict misperceptions between citizens and foreigners across the globe. *PNAS Nexus* **1**, 1–9 (2022).
9. H. Oosterbeek, R. Sloof, G. van de Kuilen, Cultural differences in ultimatum experiments: Evidence from a meta-analysis. *Exp. Econ.* **7**, 171–188 (2004).
10. E. Fehr, K. M. Schmidt, A theory of fairness, competition, and cooperation. *Q. J. Econ.* **114**, 817–868 (1999).
11. A. G. Sanfey, J. K. Rilling, J. A. Aronson, L. E. Nystrom, J. D. Cohen, The neural basis of economic decision-making in the ultimatum game. *Science* **300**, 1755–1758 (2003).
12. S. T. Fiske, A. J. C. Cuddy, P. Glick, Universal dimensions of social cognition: Warmth and competence. *Trends Cogn. Sci.* **11**, 77–83 (2007).
13. K. E. G. Williams, O. Sng, S. L. Neuberg, Ecology-driven stereotypes override race stereotypes. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 310–315 (2016).
14. M. Devaine, J. Daunizeau, Learning about and from others' prudence, impatience or laziness: The computational bases of attitude alignment. *PLoS Comput. Biol.* **13**, 1–28 (2017).
15. O. FeldmanHall, A. Shenhav, Resolving uncertainty in a social world. *Nat. Hum. Behav.* (2019), 10.1038/s41562-019-0590-x.
16. M. J. J. Handgraaf, E. Van Dijk, R. C. Vermunt, H. A. M. Wilke, C. K. W. De Dreu, Less power or powerless? Egocentric empathy gaps and the irony of having little versus no power in social decision making. *J. Pers. Soc. Psychol.* **95**, 1136–1149 (2008).
17. E. Pulcu, M. Haruno, Value computations underlying human proposer behavior in the ultimatum game. *J. Exp. Psychol. Gen.* **149**, 125–137 (2020).
18. J. Daunizeau, V. Adam, L. Rigoux, VBA: A probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS Comput. Biol.* (2014), 10.1371/journal.pcbi.1003441.
19. S. Palminteri, V. Wyart, E. Koechlin, The importance of falsification in computational cognitive modeling. *Trends Cogn. Sci.* **21**, 425–433 (2017).
20. S. Palminteri, M. Pessiglione, Opponent brain systems for reward and punishment learning: Causal evidence from drug and lesion studies in humans. *Decis. Neurosci. An Integr. Perspect.* **291–303** (2017).
21. R. C. Wilson, Y. Niv, Is model fitting necessary for model-based fMRI? *PLoS Comput. Biol.* **11**, 1–21 (2015).
22. N. Kolling *et al.*, Value, search, persistence and model updating in anterior cingulate cortex. *Nat. Neurosci.* **19**, 1280–1285 (2016).
23. A. Olsson, E. Knapska, B. Lindström, The neural and computational systems of social learning. *Nat. Rev. Neurosci.* **21**, 197–212 (2020).
24. M. Cikara, E. Bruneau, J. J. Van Bavel, R. Saxe, Their pain gives us pleasure: How intergroup dynamics shape empathic failures and counter-empathic responses. *J. Exp. Soc. Psychol.* **55**, 110–125 (2014).
25. Y. Dunham, Mere membership. *Trends Cogn. Sci.* **22**, 780–793 (2018).
26. A. Pedroni *et al.*, The risk elicitation puzzle. *Nat. Hum. Behav.* **1**, 803–809 (2017).
27. R. Frey, A. Pedroni, R. Mata, J. Rieskamp, R. Hertwig, Risk preference shares the psychometric structure of major psychological traits. *Sci. Adv.* **3**, 1–14 (2017).
28. L. J. Chang, A. Smith, M. Dufwenberg, A. G. Sanfey, Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron* **70**, 560–572 (2011).
29. T. Singer, H. D. Critchley, K. Preuschoff, A common role of insula in feelings, empathy and uncertainty. *Trends Cogn. Sci.* **13**, 334–340 (2009).
30. G. Bellucci, C. Feng, J. Camilleri, S. B. Eickhoff, F. Krueger, The role of the anterior insula in social norm compliance and enforcement: Evidence from coordinate-based and functional connectivity meta-analyses. *Neurosci. Biobehav. Rev.* **92**, 378–389 (2018).
31. M. A. J. Apps, M. F. S. Rushworth, S. W. C. Chang, The anterior cingulate gyrus and social cognition: Tracking the motivation of others. *Neuron* **90**, 692–707 (2016).
32. T. E. J. Behrens, L. T. Hunt, M. F. S. Rushworth, The computation of social behavior. *Science* **324**, 1160–1164 (2009).
33. M. Schurz *et al.*, Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychol. Bull.* **147**, 293–327 (2021).
34. N. D. Daw, Trial-by-trial data analysis using computational models. *Decis. Making, Affect. Learn. Atten. Perform.* **XXIII**, 1–26 (2011).
35. R. C. Wilson, A. G. Collins, Ten simple rules for the computational modeling of behavioral data. *Elife* **8**, e49547 (2019).
36. M. Lebreton *et al.*, Two sides of the same coin: Monetary incentives concurrently improve and bias confidence judgments. *Sci. Adv.* **4**, 1–14 (2018).
37. K. H. Schlag, J. Tremewan, J. J. van der Weele, A penny for your thoughts: a survey of methods for eliciting beliefs. *Exp. Econ.* **18**, 457–490 (2015).
38. M. Rojek-Giffin, M. Lebreton, Learning rules of engagement for social exchange within and between groups. Open Science Framework. https://osf.io/rkbev/. Deposited 21 April 2023.
39. M. Rojek-Giffin *et al.*, Neuroimaging dataset for "Learning rules of engagement for social exchange within and between groups." OpenNEURO. https://openneuro.org/datasets/ds004553/versions/1.0.1. Deposited 19 April 2023.