Rational Social Valuation Inference in Humans: Supplementary Online Materials

July 6, 2022

Contents

1 The Welfare-Tradeoff task **3**

2 The model **4**

3 Parametrization of the model **5**

3.1 Noise parameter 6

3.2 First parametrization of the prior 6

3.3 Second parametrization of the prior 7

4 Individual-level analyses **9**

4.1 Correlations between model predictions and participant predictions 9

4.2 Association between Inferred WTR and participant predictions 10

4.3 Association between Inferred WTRs and emotion 12

5 Analysis of study 1 with the first prior parametrization **15**

6 Analysis of study 2 with the first prior parametrization 17

7 Deviations from pre-registration 18

References 21

1 The Welfare-Tradeoff task

The Welfare-Tradeoff Task is adapted from Delton (2010). This task involves a "dictator" and a "recipient". In each trial, the dictator has to choose between allocating a monetary payoff of $\pi_{dictator}$ to himself, or a payoff of $\pi_{recipient}$ to the recipient (henceforth, we refer to allocating the money to the recipient as 'Giving', and allocating the money to oneself as 'Taking'. Across trials, the value of $\pi_{dictator}$ varies, but the value of $\pi_{recipient}$ remains constant or varies only minimally. The recipient does not make decisions).

Here we ask: can people, after observing a few decisions the dictator made in this task, predict his behavior in other trials? This is a problem of inference under uncertainty: observing a few decisions gives the observer some information about the dictator's psychology, but not enough to achieve complete predictive power.

To develop an ideal observer model of this inference process, we first need to make some assumptions about the behavior of dictators in this task. We assume that each dictator has a "Welfare-Tradeoff Ratio" (WTR) for the task. The behavior of the dictator in a given trial is governed by the following simple decision rule:

"If $WTR > \frac{\pi_{dictator}}{\pi_{recipient}}$, allocate the money to the recipient; otherwise allocate the money to yourself."

Existing empirical work (Delton, 2010) suggests that this decision rule is a plausible model for behavior on the Welfare Tradeoff Task.

Empirical work has also found that dictators are not always perfectly consistent with their WTR. To account for this, we assume that dictators' observation of a trial is noisy. For each trial, the ratio between the potential benefit to the dictator and the potential benefit to the recipient is denoted by $\phi = \frac{\pi_{dictator}}{\pi_{recipient}}$. We assume that the dictator observes this value with some noise, such that he observes $\phi + \epsilon$, where ϵ is drawn from a normal distribution with mean 0 and variance σ_{ϕ}^2. Then the dictator allocates the money to the recipient ("Gives") if $WTR > \phi + \epsilon$.

2 The model

The ideal observer model takes as inputs choices made by a dictator, and uses them to update its belief about the dictator's WTR. When observing a trial, the ideal observer updates its belief about the WTR of the dictator according to Bayes' rule:

$$p(WTR|\phi, decision) = \frac{p(decision|\phi, WTR)p(WTR)}{p(decision|\phi)}$$

"decision" can take two possible values "Give" or "Take". $p(WTR)$ is the model's prior about the dictator's WTR (see next section for information on how we set its initial value).

The likelihood $p(decision|\phi, WTR)$ is defined by the following equations, depending on whether the dictator chooses to Give or Take:

$$p(\text{``}Give\text{''}|\phi, WTR) = p(WTR > \phi + \epsilon)$$

$$p(\text{“}Take\text{”}|\phi, WTR) = p(WTR < \phi + \epsilon)$$

where $\epsilon \sim \mathcal{N}(0, \sigma_\phi^2)$.

The denominator $p(decision|\phi) = \int p(decision|\phi, WTR)p(WTR)dWTR$ is a scaling factor. We set $\sigma_\phi = .16$ (see section 3 below).

The model can then be used to predict the probability that a dictator will Give in a trial with ratio ϕ.

$$p(\text{“}Give\text{”}|\phi) = \int p(\text{“}Give\text{”}|\phi, WTR)p(WTR|*)dWTR$$

where $p(WTR|*)$ is the belief that the model has about the WTR of the dictator (here “ $*$ ” denotes the observations that the model has made prior to that judgment). The likelihood function $p(\text{“}Give\text{”}|\phi, WTR)$ is defined, as above, as $p(\text{“}Give\text{”}|\phi, WTR) = p(WTR > \phi + \epsilon)$, with $\epsilon \sim \mathcal{N}(0, \sigma_\phi^2)$.

Computations were made in R, using grid approximation.

3 Parametrization of the model

The ideal observer model needs to be equipped with a prior distribution of WTR (which represents the observer's best guess about the dictator's WTR in the absence of any information), as well as a noise parameter σ_ϕ, which represents the amount of noise that the observer assumes goes into the dictator's choices.

3.1 Noise parameter

We have set the value of σ_ϕ on the basis of empirical data. We used data previously collected for a larger study (Sznycer et al., unpublished data) where participants (N=479, recruited on MTurk, 10 additional participants excluded for failing an attention check) played several rounds of the Welfare Trade-off Task as dictators. Here, we only analyzed trials where $\pi_{recipient} \approx \$31$ and the participant was told to imagine making trade-offs between his/her own welfare and that of a hypothetical acquaintance. We therefore computed the distribution of WTRs in the sample for the Welfare-Tradeoff task defined by $\pi_{recipient} \approx \$31$. For each participant, we computed a WTR and a Consistency score using the algorithms developed in Delton (2010, pp. 49-51).

To estimate the value of σ_ϕ, we assumed that every participant has his own value of σ_ϕ, and that the variable is distributed in the population according to a gamma distribution. Using Maximum Likelihood estimation, the distribution of Consistency scores in the sample was most consistent with the distribution of σ_ϕ in the sample following a gamma density function with $\alpha = .59$ and $\beta = 1.90$. The present ideal observer model does not attempt to infer the idiosyncratic value of σ_ϕ for every individual dictator, instead it assumes the same constant value for each dictator. Therefore we set σ_ϕ to be the median of the gamma density function with $\alpha = .59$ and $\beta = 1.90$, which yielded a value of $\sigma_\phi = .16$.

3.2 First parametrization of the prior

One potential way of determining the prior of the ideal observer is by measuring the actual distribution of WTRs in a subset of the population. Therefore, we set the first parametrization of the prior as an approximation of the distribution of WTRs in the sample we describe in the section above. The distribution of WTRs in this sample (for the task with $\pi_{recipient} \approx \$31$) was best approximated (using Maximum Likelihood estimation) by a skewed Laplace distribution with location = 0, dispersion = .23, skew = .63. We use this skewed Laplace distribution as the prior p(WTR) for the ideal observer.

3.3 Second parametrization of the prior

One potential downside of the first parametrization is that we estimated the distribution of WTRs in MTurk participants, which may not exactly match the population that comes to the mind of our participants when they think of their acquaintances. Another potential concern is that the Welfare-Tradeoff Task was probably new to most of our participants, and even if we assume that they have a good generative model of how people typically behave when they make welfare trade-offs, there is no strong reason to expect them to have perfectly accurate priors for that specific task. Therefore, for the second parametrization of the prior, we directly attempted to infer the prior belief that participants had about the distribution of WTRs among their acquaintances.

We did so by asking participants in study 2 to complete a preliminary task, at the beginning of the experiment, where they had to make predictions about the behavior of interaction partners in the WTT for which they had not had an opportunity to see any other WTT decision (see 'methods' section in the main text for details). For this part of the analysis, we only analyzed data from participants who exhibited a negative correlation between the cost of giving in a trial ($\pi_{dictator}$) and the participant's prediction for that trial during this preliminary prediction task. 8 participants did not meet that criterion, and thus did not provide interpretable data.

We first ran a multilevel quadratic regression on this data, with the participants' predictions as an outcome variable, and cost of giving in a trial as a predictor variable, with intercepts and slopes (for both the first and second-order term of the polynomial) varying across participants. Using the coefficients from that model, for each participant we generated simulated predictions for each trial of this preliminary prediction task.

We assumed that the prior of a participant i about the distribution of WTRs among his/her acquaintances follows a normal distribution with mean μ_i and standard deviation σ_i. One can compare the simulated predictions for a participant with the predictions made by an ideal observer with a given prior. By systematically varying the parameters μ_i and σ_i in the prior used by the ideal observer, one can find a best-fitting pair (μ_i, σ_i) for each participant i, using least squares optimization. Using this approach, across participants, we estimated an average $\hat{\mu}$ of .55 (95%CI: .41 - .69), and an average $\hat{\sigma}$ of 1.01 (95%CI: .89, 1.13). We used these parameters for the prior of the ideal observer. That is, the prior of the ideal observer for a partner's WTR is a normal distribution with $\mu = .55$, $\sigma = 1.01$.

4 Individual-level analyses

In the main text, we report item-level correlations, which collapse across participants, and linear mixed models, which partially pool data across participants. Because each participant made 50 predictions and made 20 emotion ratings, we can also treat each participant as its own statistical universe, and perform analyses at the individual level. Doing so provides a robustness check, ensuring that results reported in the main text are not an artifact of averaging. Here we use boxplot graphs to report the results of 650 statistical models, each performed on data from one participant.

4.1 Correlations between model predictions and participant predictions

For each participant, we computed the correlation, across trials, between the participant's predictions and the ideal observer predictions. Figure S1 reports the distribution of these correlation coefficients, showing that for most participants, there was a close fit between model and participant predictions.

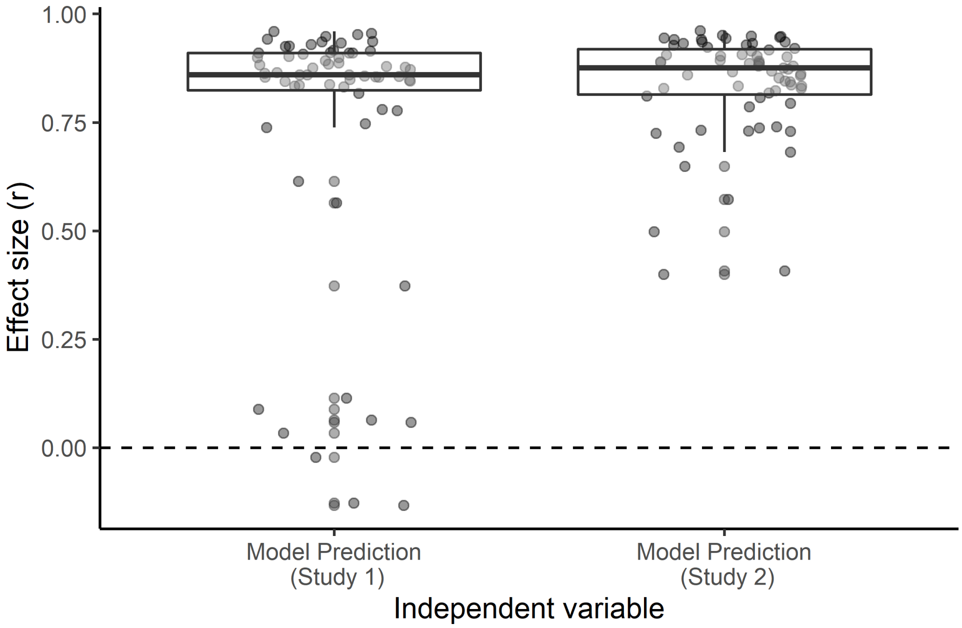

Figure S1: Pearson's correlation coefficients (r) for the association between model predictions and participant predictions. Each data point corresponds to one correlation coefficient (i.e. to one participant). Points are jittered along the x-axis for readability.

4.2 Association between Inferred WTR and participant predictions

Do participant predictions reflect inferences about social valuation? If so, we would observe a positive correlation between a participant's predictions about a partner and the WTR inferred by the ideal observer for that partner. This association should still hold, even controlling for material payoffs.

For each participant, we computed two linear regression models, with inferred WTR as an IV and participant prediction as a DV. The second model also had material payoffs as an additional IV. For each test, we extracted the standardized regression coefficient for the inferred WTR variable. Figure S2 reports the distribution of these coefficients, showing that for most participants, there was a close fit between model and participant predictions, and this association subsisted, although it was attenuated, when controlling for material payoffs.

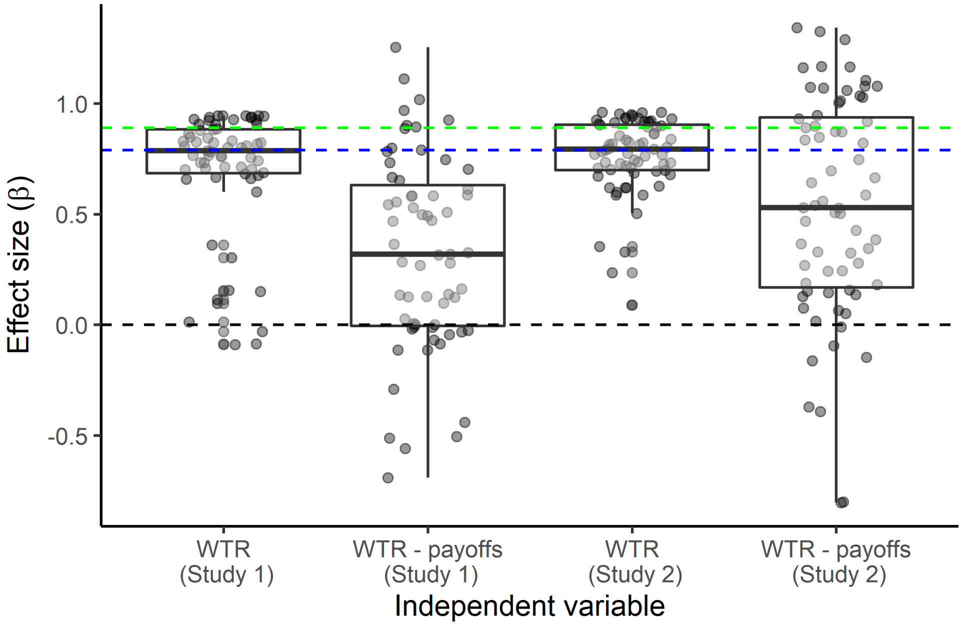

Figure S2: Standardized regression coefficients (β) for the association between ideal-observer-inferred WTR and participant predictions. Each data point corresponds to one coefficient (i.e. to one participant). 'WTR': zero-order association between inferred WTR and prediction; 'WTR – payoffs': association between inferred WTR and predictions, controlling for material payoffs. The green dashed line corresponds to the association between ideal-observer-inferred WTR and the model predictions. The blue dashed line corresponds to the same value, controlling for material payoffs. Points are jittered along the x-axis for readability.

4.3 Association between Inferred WTRs and emotion

For each participant, we computed four linear regression models, two for Anger and two for Gratitude. For each emotion, the first model had Inferred WTR as an IV and Anger (or Gratitude) as a DV. The second model also had material payoffs as an additional IV. For each test, we extracted the standardized regression coefficient for the inferred WTR variable. Figures S3–S4 report the distribution of these coefficients. For most participants, the WTR inferred by the ideal observer for a partner was a strong predictor of the participant's Anger and Gratitude toward that partner. For Gratitude, this association considerably weakened when controlling for material payoffs. For Anger, this association remained (on average) unchanged even when controlling for material payoffs.

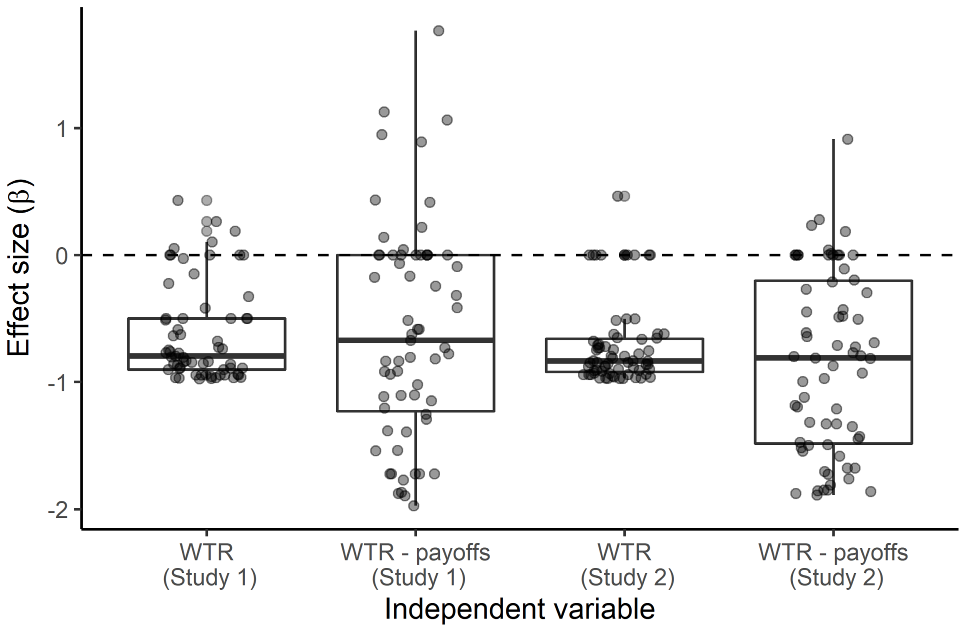

Figure S3: Standardized regression coefficients (β) for the association between ideal-observer-inferred WTR and participant Anger. Each data point corresponds to one coefficient (i.e. to one participant). ‘WTR’: zero-order association between inferred WTR and Anger; ‘WTR – payoffs’: association between inferred WTR and Anger, controlling for material payoffs. Points are jittered along the x-axis for readability.

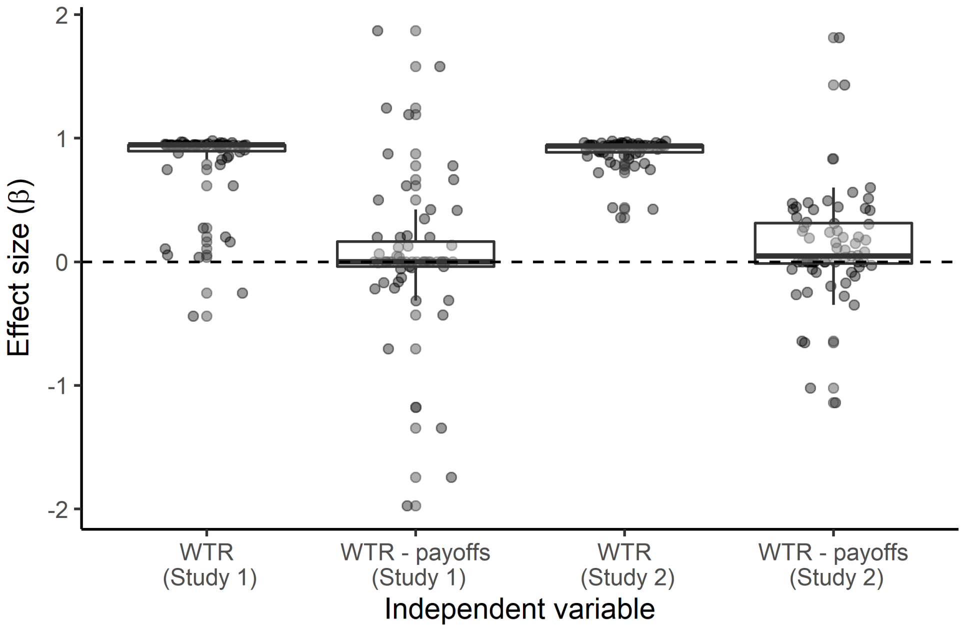

Figure S4: Standardized regression coefficients (β) for the association between ideal-observer-inferred WTR and participant Gratitude. Each data point corresponds to one coefficient (i.e. to one participant). 'WTR': zero-order association between inferred WTR and Gratitude; 'WTR – payoffs': association between inferred WTR and Gratitude, controlling for material payoffs. Points are jittered along the x-axis for readability.

5 Analysis of study 1 with the first prior parametrization

Do human predictions match ideal observer predictions?

Yes. The item-level correlation between the average human prediction for a given trial and the model prediction for that trial was $r(48) = .86$, $p < .001$.

Human predictions also correlated with model predictions when analyzed at the individual level: the median correlation between an individual's predictions and the model predictions (across trials) was $r(48) = .75$; inter-quartile range: .66 to .84.

Can this result be entirely explained by a simple heuristic which tracks material payoffs?

No. A hierarchical linear model with partner's type and inferred WTR as predictors, random slopes and random intercepts, and participant as a random effect, shows that, controlling for material payoffs, the WTR inferred by the ideal observer remains positively associated with human predictions, $b = .09$, $p = .006$, suggesting that participants did make social valuation inferences.

Does inferred WTR predict anger and gratitude?

Yes for Anger, no for Gratitude. Hierarchical linear models with inferred WTR as predictors, random slopes and random intercepts, and participant as a random effect, show that inferred WTR is a negative predictor of Anger, $b = -.47$, $p < .001$, and a positive predictor of Gratitude, $b = .76$, $p < .001$.

However, when controlling for material payoffs, inferred WTR was no longer a significant predictor of Gratitude, $b = -.005$, $p = .82$, though it remained a significant predictor of Anger, $b = -.08$, $p = .02$.

6 Analysis of study 2 with the first prior parametrization

Do human predictions match ideal observer predictions?

Yes. The item-level correlation between the average human prediction for a given trial and the model prediction for that trial was $r(48) = .877$, $p < .001$. Human predictions also correlated with model predictions when analyzed at the individual level: the median correlation between an individual's predictions and the model predictions (across trials) was $r(48) = .78$; inter-quartile range: .66 to .84.

Can this result be entirely explained by a simple heuristic which tracks material payoffs?

No. A hierarchical linear model with material payoffs and inferred WTR as predictors, random slopes and random intercepts, and participant as a random effect, shows that, controlling for material payoffs, the WTR inferred by the ideal observer remains positively associated with human predictions, $b = .16$, $p < .001$, suggesting that participants did make social valuation inferences.

Does inferred WTR predict anger and gratitude?

Yes. Hierarchical linear models with inferred WTR as predictors, random slopes and random intercepts, and participant as a random effect, show that inferred WTR is a negative predictor of Anger, $b = -.56$, $p < .001$, and a positive predictor of Gratitude, $b = .81$, $p < .001$.

When controlling for material payoffs, inferred WTR remained a significant predictor of Gratitude, $b = .06$, $p = .02$, and Anger, $b = -.13$, $p = .005$.

7 Deviations from pre-registration

Here we address three aspects in which the contents of the current paper deviate from the pre-registration.

Are individual differences in emotion related to individual differences in predictions?

In our pre-registrations, we made the hypothesis that participant predictions would be correlated with their emotion ratings, even when holding the stimuli constant. That is, for participants observing the same partner, participants reporting higher anger (lower gratitude) would subsequently predict a lower likelihood that the partner will allocate the money to the participant in observation trials.

We found some support for the hypothesis. However, it occurred to us that the data might be explained by the following confound. Consider two participants: participant A pays little attention to the task and responds randomly, while participant B pays close attention. When evaluating a selfish partner, participant B will make less optimistic predictions, and will also report higher anger, than participant A. Therefore, the presence of many inattentive participants would on its own be enough to cause a correlation between predictions and emotion ratings.

In order to control for this possibility, we computed, for each participant, the correlation between that participants' predictions and the ideal observer predictions across trials. We used this variable as a proxy for the participants' engagement with the task. We found that, controlling for engagement with the task, there was no correlation between a participant's predictions and their ratings of gratitude, in either study 1 or study 2. In both study 1 and 2 we found a correlation between predictions and anger ratings, but the associated p-values were not very far below the conventional threshold for statistical significance. Therefore, the current data do not speak strongly either in favor or against the hypothesis. This is why we choose the report the analyses here instead of the main text.

Study 1

We ran linear mixed models with random slopes and random intercepts, participant prediction as fixed effect, and partner identity as a random effect. We find that participant predictions are negatively associated with Anger, $b = -.17$, $p = .009$, and positively associated with Gratitude, $b = .23$, $p = .001$. Running the same analyses with engagement with the task as an additional covariate, we find that predictions are negatively associated with Anger, $b = -.08$, $p = .03$, but were not associated with Gratitude, $b = .04$, $p = .12$.

Study 2

We ran linear mixed models with random slopes and random intercepts, participant prediction as fixed effect, and partner identity as a random effect. We find that participant predictions are negatively associated with Anger, $b = -.09$, $p = .02$, but are not associated with Gratitude, $b = .004$, $p = .84$. Running the same analyses with engagement with the task as an additional covariate, we find that predictions are negatively associated with Anger, $b = -.07$, $p = .04$, but were not associated with Gratitude, $b = .004$, $p = .84$.

Letters designating the partners

For convenience, we would want the alphabetical order of the letters by which we refer to interaction partners to be the same as the ordering of the WTR they express toward the participant – but this was not the case in the pre-registrations. Therefore, we changed the letters in order to fix this issue. As a result, the letters used in the paper do not match those in the pre-registration. This is just a cosmetic issue and has no implications for any of the analyses we run. Note that partners were never identified by their letters to participants.

Classification of participants

In the pre-registration, we planned to analyze separately the participants who showed a significant negative correlation between the ratio $\phi = \pi_{partner}/\pi_{participant}$ on a prediction trial and the participant's prediction for that trial. Looking back, this criterion is too harsh as a test of whether a participant paid attention to the task. Even the ideal observer predictions are correlated with ϕ at only $r = -.40$, and with 50 trials per participants, any correlation with $|r| < .27$ fails to be significant at the conventional .05 level. Therefore, we eventually decided against discriminating participants on this basis. Analyzing only participants who meet the criterion does not fundamentally alter the studies' findings: in both study 1 (N=33) and study 2 (N=41), the item-level correlation between average human prediction and ideal observer prediction is $r(48) = .98$, $p < .001$.

References

Delton, A. W. (2010). *A psychological calculus for welfare tradeoffs.* University of California, Santa Barbara.