www.sciencemag.org/content/361/6401/eaat0036/suppl/DC1

Supplementary Materials for

Computational and neurobiological foundations of leadership decisions

Micah G. Edelson*, Rafael Polania, Christian C. Ruff, Ernst Fehr*, Todd A. Hare*

*Corresponding author. Email: micah.edelson@econ.uzh.ch (M.G.E.); todd.hare@econ.uzh.ch (T.A.H.); ernst.fehr@econ.uzh.ch (E.F.)

Published 3 August 2018, *Science* **361**, eaat0036 (2018)
DOI: 10.1126/science.aat0036

This PDF file includes:

Materials and Methods
Supplementary Results
Figs. S1 to S8
Tables S1 to S7
References
Appendices S1 and S2

TABLE OF CONTENTS:

I. Materials and Methods.

1. Participants and sample size determination.

We conducted the experiment with two separate samples of participants – marked throughout the manuscript as original and fMRI replication groups. The difference between the groups was that the latter performed the delegation task in the MRI scanner. Previous laboratory experiments on individual versus group decision making have typically used between 30-50 participants (*37–39*). Power calculations (*40*) based on the aforementioned studies average effect sizes suggested a stopping criterion of 40 participants as a reasonable estimate to ensure a statistical power of 0.8 (with an alpha level of 0.05). We thus recruited 40 participants for the original group (21 females; age 25.7 ± 0.66 standard error of the mean). In the fMRI replication group, we added, *a priori*, four additional participants (constituting one unit of participants, see below, resulting in 44 participants; 25 females; age 23.5 ± 0.43). This was done in anticipation of some minor data loss due to issues such as excessive head movement in the scanner, and because the minimum experimental session size could not be under eight participants (see task design below). The data for three participants were not fully collected (two participants failed the test quiz assessing comprehension of the instructions and one participant did not show up for the second stage), resulting in a final N=38 and N=43 for the original and fMRI replication groups respectively. All participants gave informed consent and were remunerated for their participation. The study was approved by the Ethics Committee of the Canton of Zurich.

In the original experiment, Stage1 (see Methods 2.1 below) started with 20 participants randomly assigned to five groups, each consisting of four unrelated individuals. Blind randomization was performed by asking individuals to choose among a shuffled stack of identical looking cards with concealed labels. For the fMRI replication experiment, given that using a functional magnetic resonance imaging (fMRI) scanner limits the potential number of participants that can be measured in a day, the number of individuals participating in each Stage 1 session was reduced. The size of each group remained the same, but three Stage1 sessions consisted of three groups each (i.e., 12 participants per round) and a final Stage 1 session consisted of two groups (i.e., 8 participants). This allowed all fMRI replication group participants from a given Stage 1 session to complete Stage 2 (see 2.2 below) of the experiment within a maximum of four-days from one another.

2. Task design and self-report measures.

Both groups participation in the experiment involved 2 stages. In stage 1 they performed the *Baseline task* while in stage 2, which took place two days later in the original group and three to six days later in the fMRI replication group, they conducted the *delegation task*.

2.1 Stage 1

2.1.1 Group induction phase. In order to form a sense of group coherence within each set of four previously unacquainted individuals, stage 1 started with a group induction procedure. The procedure followed standard group induction protocols (*7*). Individuals in each group received a colored ID tag identifying their group. Participants were informed that they would perform several quizzes as a group and that their group performance would be compared to the other groups in that experimental session. The best performing group earned a bonus of 60 CHF (~55 €). During the group induction phase, each group was seated together and was given 15 minutes to jointly answer a quiz consisting of music-related questions. Following this quiz, each group was divided into two pairs. Each individual in the pair was given three minutes to describe themselves to their partner in as much detail as possible. Participants were informed they would be tested about this information later. Following this stage, each pair was given 10 minutes to answer a quiz containing 20 general questions related to basic history. The pairs were then changed and the procedure repeated again (including the personal description and a quiz, this time related to art). The aggregate performance of the group on all quizzes determined the winning group who received a prize of 60 CHF at the end of the experiment. In order to avoid the possibility that the outcome of this stage will influence the rest of the experiment, the identity of the winning group was not revealed to the participants until the whole experiment ended. This type of group induction procedure is commonly used (*7*) to establish a minimal level of acquaintance within a group of individuals who were ex-ante strangers.

2.1.2 Baseline task (Fig. 1A). After the completion of the group-induction phase, each participant was seated in a separate cubicle and performed the baseline task independently. Decisions on this task were not related to the other members of the group. Participants were faced with 200 decisions.

On each trial participants had to choose whether or not to take a risky action. Each risky action was associated with a probability of success and failure (proportion of green and red wedges respectively) depicted on the screen as colored slices of a 10-piece probability pie. In order to eliminate the necessity for counting the slices, these probabilities were also depicted in adjacent text (**Fig 1A**). The potential reward if the gamble was successful, or loss if it failed, were also presented on the screen. A decision not to take the risky action always resulted in a sure outcome of 0. In order to increase engagement in the task, the participants were told to imagine themselves lost in a jungle. Each decision was framed in the context of the possible action a stranded person (or group) could take (e.g., cross a river, light a fire). The question frames were randomized across the different questions for each participant and did not affect the results. The question order was randomly assigned for each participant.

In real life circumstances, the true underlying probabilities of success are almost never perfectly known to the decision maker before she acts. Thus, to emulate ecologically realistic situations, we added an element of uncertainty. On 140 of the 200 trials, a gray cover of varying size (see range of stimuli below) obscured part of the probability circle. The participants were told that beneath this cover could be any proportion of red or green slices, and they must make a decision based on this partial information. The inclusion of experimentally controlled uncertainty allowed us to additionally test theoretical predictions concerning the relationship between efficiency, ambiguity preference, responsibility aversion and leadership (*1*, *41*, *42*). Participants did not receive any feedback on the outcome of their choices at this stage of the experiment.

2.1.3 Range of stimulus values and payments. The portion of the circle covered by the gray area ranged in size between 1-9 slices, with a uniform distribution across slices. On 60 additional trials, no cover was presented, i.e., these were the pure risk trials with perfect information about the probabilities. The possible gains and losses ranged from +10 to +100 points and -10 to -90 points respectively. The probability of success ranged from 10% up to 90%. The specific combinations of gain, loss, probabilities and cover size were pseudo-randomly chosen to result in a normally distributed expected-value distribution that maximized the degree of orthogonality between the different components of the expected value while maintaining the aggregate informational advantage that played a key role in the next stage of the protocol (Delegation task, see below). The expected payment distribution had a positive mean (18.4 points) calculated to provide participants with average earnings of 25 CHF per hour when including all payments across both stages of the experiment.

For earnings at stage 1, five trials were randomly selected and the sum of earned points on these trials was converted to CHF (with a conversion factor of 0.4). This procedure ensured that participants would need to perform well on every trial regardless of their performance on previous trials. Note that payment for all stages of the protocol (including the group quiz and baseline task) was performed at the end of stage 2 and participants were not exposed to feedback on their performance or earnings at stage 1.

2.1.4 Ambiguity preference test. After performing the baseline task, participants completed the a modified Ellsberg ambiguity preference test (*43*, *44*). In successive decisions, individuals were asked to choose whether they preferred a sampler drawn from an urn with a known distribution of winning and losing balls (with progressively worse odds of success), or from an urn with an unknown distribution. The point at which the individual's preference switches between the unambiguous and ambiguous urns has been consistently demonstrated to correlate with their ambiguity preferences (*45*).

2.2 Stage 2

The participants returned to the lab two to six days after they participated in Stage 1. The participants were seated in individual cubicles (for the original group) or in a single-participant experimental room outside the fMRI scanner (for the fMRI replication group) and were instructed to perform a written memory test. In this test, participants were asked to recall all the information they remembered concerning the two other group members who provided details about themselves during the previous stage. They were also asked to re-answer the music quiz according to their memory of what the group answered in the previous stage. The objective of this memory test was to serve as a reminder of the group interaction from the previous stage.

2.2.1 Delegation task (Fig. 1B). After performing the memory task, the participants were given written instruction for the Delegation task. In this task, in addition to the option to accept or reject the risky action, participants could also defer, i.e., they could give their right to choose the risky or safe option to the other group members. If they chose to defer, the majority answer from the other three group members given during the *baseline task* for the same risky choice would be implemented for this trial. We deliberately did not include the leader's own answer in determining the majority's decision, so that deferral meant completely relinquishing decision power, in order not to induce a sense of diffused responsibility on these trials.

Before participants made decisions in the Delegation task they received detailed instructions that explained the task. They were, in particular, informed that each of the other group members saw a different part of the probability pie but that each participant faced the same amount of uncertainty, i.e., the size of covered area of the probability pie was identical across subjects in each given trial. Participants thus knew that other group members' collective information about the probability pie typically was superior to their own information but that this informational advantage varied with the amount of uncertainty/ambiguity present in the trial (see **Fig. S1**). The participants received unlimited time to read the instructions and were subsequently required to perform a three-question quiz testing their understanding. Two participants answered the majority of the questions incorrectly on this quiz and their data were excluded from our analyses (see participants section above).

The Delegation task consisted of the same 140 ambiguous trials from the *baseline task* repeated under two conditions (280 trials in total) as follows. The only difference between the matched Group and Self trials was that in the Group condition the outcome of the action affected the other group members as well as the target participant. For example, if the participant decided to gamble on a Group trial and was successful in obtaining a reward of 50 points, this amount was added to the payment of each of the four group members separately. In contrast, on Self trials, the participant's action only affected his or her own monetary payoff. The matched Group and Self trials were identical in all other respects, including the probabilities, rewards, the amount of ambiguity and the informational advantage of deferring the decision to the other group members. The question order was randomized for each participant. Group and Self trials were presented in blocks of 10 trials that were pseudo-randomly intermixed for each participant such that no more than three blocks of a given condition were presented consecutively. Given the large number of trials, it is unlikely that participants remembered specific parameter combinations associated with the matching trials across conditions. To further prompt the independent treatment of each trial, the entire probability circle was randomly rotated for matching trials across the three presentations (baseline task, Self and Group conditions in the Delegation task). Thus, although the information related to each question remained identical, the visual display of the probability pie was changed to help ensure every trial was considered independently.

In effect, each of the four participants in a group could act as leaders to directly determine the outcome of their entire group on Group trials. During the Delegation task there was no interaction between the individuals and participants could not influence the other group members' decisions. In order to minimize the possible implicit expectance of reciprocity, participants were informed this was to be the final group-related task. Moreover, in order to enhance personal accountability, participants were also informed that after termination of the experiment the amount of points they earned for the group in the Delegation task would be announced in front of their group.

Participants were given unlimited time to answer each of the 280 trials in this test (for additional RT data see Supplementary 7). After they made a choice in a trial, participants received feedback regarding the outcome (i.e. amount gained or lost) of their choice and/or the outcome of the group's majority choice which was displayed on the screen for 2 seconds. Participants were always shown the outcome that would have resulted from the group's majority choice, regardless of whether or not they deferred on that trial. The outcome for their own choice was only shown if they opted to make a choice themselves on that trial. Note that the feedback was identical for the Self and Group trials, and thus cannot explain differences in deferral behavior between conditions (nonetheless, see Supplementary 5 for validation experiment without feedback).

For improved temporal separation between conditions in fMRI imaging, in the fMRI replication group fixation crosses with a pseudo-random duration (mean 3.8, s.d 1.7) were added after the participant's decision and after feedback presentation. The durations of the fixations were optimized for our specific task using the behavioral data from the original group and were randomly allocated across trials for each individual.

After completing the Delegation task, participants filled in the leadership measures detailed in the next section.

2.3 Leadership measures collected at the end of Stage 2

Although there are different categorizations of leadership (*25*, *46–60*), the majority of classifications systems include some aspects of Goal-oriented leadership which emphasizes the accomplishment of task objectives (*11*, *12*). Therefore, the current task was specifically designed to assess the goal-oriented aspect of leadership. The relationship oriented aspect of leadership which emphasizes behaviors that facilitate long-term team development and inter-personal interaction is less relevant in our protocol since participants cannot interact directly or influence the behavior of others in the delegation task. For parsimony and robustness, we assessed goal oriented leadership by means of two of the most widely used measures directly targeting this aspect of leadership (*1*, *11*, *12*). The original group participants completed the Leadership Behavioral Description Questionnaire (LBDQ). The fMRI replication group participants completed the LBDQ as well as the Blake-Mouton Managerial Grid Questionnaire (BMMG).

Here, we deliberately used simple and basic leadership measures to capture core aspects of leadership (*1*, *41*). It would also be informative to link responsibility aversion to individual leadership concepts in the future [e.g., Transformational Leadership (*21*, *22*, *57*), Destructive leadership (*23*), the role of followers (*26*, *55*, *61*), situational factors (*62*), gender differences (*56*), and additional personality traits (*46*, *57*)]. Three participants had missing values in the questionnaires and therefore could not be included in leadership-related analyses.

2.3.1 Leadership Behavioral Description Questionnaire (LBDQ) (*9*)**.** This questionnaire is a validated measures of leadership ability (*11–13*). It consists of two independent sub-scales measuring the goal-oriented and relationship oriented leadership aspects mentioned above. The scores on both these sub-scales have been repeatedly related to real-life leadership positions and ability in numerous fields including business, politics and sports over the last 50 years (*1*, *11*, *12*, *63*).

2.3.2 Blake-Mouton Managerial Grid (BMMG) (*10*)**.** This measure explores leadership attitudes rather than behaviors and forms the basis for commonly employed management training programs in business firms (*64, 65*). This questionnaire also includes two sub-scales corresponding to goal oriented and relationship oriented leadership.

Due to time constraints, we divided the original group sample into two. Half the participants were randomly assigned to fill in the LBDQ target subscale (goal oriented leadership). A highly significant correlation was found between responsibility aversion and the goal-oriented component (*rho*=-0.77, *p*=2*10^{-4}). The other half of the participants performed the relationship oriented leadership subscale. Although we also found a significant correlation here, it did not survive correction for multiple tests across the two comparisons (*p*=0.04, uncorrected). While Responsibility aversion may correlate with relationship oriented leadership given a larger sample size, we decided to focus on the goal oriented component which is also the most relevant to our task design (*66*). Subsequently, all individuals in the fMRI replication group answered the LBDQ target subscale as well as a second goal oriented leadership measure (BMMG; *"concern for results"* subscale).

2.3.3 Composite leadership score. To reduce potential biases contained in any one questionnaire, and to incorporate both sources of information in the fMRI replication group into a single measure, we used an average over the two normalized task leadership measures (Composite leadership score) as our primary measure of leadership in the fMRI replication group. For completeness, we also list below the correlation between our model's Responsibility aversion effect (i.e. the change in the deferral thresholds) and each of these measures separately. Mirroring the findings noted in the main text for the composite score, we found a negative correlation with both measures separately (*rho*=-0.34, *p*=0.03 and *rho*=-0.29, *p*=0.06 for the BMMG and LBDQ measures, respectively).

2.3.4 Real-life leadership measure. Two features of Swiss society, namely, mandatory military service for males and the wide popularity of the Swiss Scouts organization, provided us with an opportunity to test whether our responsibility aversion measure correlated with leadership behaviors beyond those captured in the self-report questionnaires. We asked each participant to provide their military rank, and for those who led groups in the Scouts, the number of years they led and the size of their groups. We then ranked the scores for each question separately, and averaged across the measures available for each participant, to create a combined score. Participants who did not participate in the army or in the scouts could not be included in this analysis, resulting in a final sample size of N=21. As reported in the main text, Responsibility aversion was the only measure that significantly correlated with these real-life expressions of leadership (**Fig. 2,** *rho* =-0.49, *p*=0.02; whereas no significant correlation was found with the preferences for decision rights in *Group* or *Self* conditions; rho=0.03 p=0.84; rho=-0.13 p=0.33, respectively).

2.4 Social preference measures

Participants also performed an anonymous dictator game task in which they received an endowment of 7 CHF (~6 €) and could decide to allocate any part of this money to a random member of their group. This procedure was repeated with a random out-group member in the same experimental session. The results obtained in this task are in line with those previously found in the literature (*67*) (average transfer 42% ±3% and 24% ±3% of the total available sum to the in-group and out-group, respectively). In addition, participants rated their feeling of affiliation with their in-group and out-groups on a 1-10 scale. The average affiliation with the in-group was significantly higher than with the out-group (5.3±0.29 vs 2.1 ±0.27; *z*=9.9, *p*≈0). One participant in the fMRI replication group who did not fill-in the in-group affiliation score could not be included in analyses that required this value.

2.5 Payment

After the termination of all stages of the protocol, participants were paid according to their accumulated performance on all experimental tasks. For the delegation task, five random trials from each of the conditions (Group/control) were selected for payment. As aforementioned, in order to enhance personal accountability, before each individual was paid, the whole group was informed about the performance of each group member on Group trials in the delegation task (but not about performance on any other part of the protocol). Participants were then separated and paid according to their total performance. We also ran an additional control experiment without this enhanced accountability feature and found that participants were still responsibility averse (see Supplementary, Section 5).

3. Computational modeling:

3.1. Prospect theory (PT) model description.

Choices in the *baseline task* were fit using cumulative prospect theory (*18*, *68*) , which is based on the assumption that the expected subjective value of the risky option u is defined by

$$u = v(x_g)\pi(p_g) - v(x_l)\pi(p_l), \qquad (1)$$

where $v(.)$ represents the value function, x_g and x_l denote the potential gains and losses, respectively, involved in the prospect, p_g and $(1 - p_g)$ denote the corresponding probabilities of a gain and a loss whereas $\pi(p_g)$ and $\pi(1 - p_g)$ are the subjective decision weights attached to these probabilities.

The value of the safe option v(0) is normalized to be zero; therefore the expected subjective value of the risky prospect as given in (1) also describes the expected subjective value difference between the risky and the safe option. The value function v(.) has the following properties:

$$v(x) = \begin{cases} x^{\alpha} & \text{if } x \geq 0 \text{ (i.e. } x_g) \\ -\lambda(-x)^{\alpha} & \text{if } x < 0 \text{ (i.e. } x_l) \end{cases}, \qquad (2)$$

where λ is the loss preference parameter and α determines the concavity of the value function (given a correlation between the α and the τ parameter in eq. 5 below, we fixed the value of α

according to previous literature at 0.9 (*18*, *69*), however for completeness we estimated the model with α ranging from 0.5-1 and obtained equivalent results). To accommodate for the existence of unknown probabilities (i.e., for ambiguity), the probability (η) by which the outcome *x* occurs is defined by

$$\eta = \frac{1}{N}\sum_{i=1}^{N} z_i, \text{ where} \begin{cases} z_i = 1 & \text{if piece of pie is green} \\ z_i = 0 & \text{if piece of pie is red} \\ z_i = \theta & \text{if piece of pie is grey} \end{cases}, \tag{3}$$

where *N*=10 is the number of slices in the stimulus pie (see **Fig. 1**), and θ represents an ambiguity preference parameter that ranges between 0 and 1. Given that the gray pie slice is equally likely to be red or green this parameter would be 0.5 for an ambiguity neutral agent. Estimates under 0.5 indicate that the agent is ambiguity averse and assigns larger chance of failure than success to the unknown part of the probability pie. Probabilities are transformed by a non-linear weighting function

$$\pi(\eta) = \frac{\eta^{\gamma}}{(\eta^{\gamma} + (1-\eta)^{\gamma})^{1/\gamma}}, \tag{4}$$

where γ specifies the s-shaped transformation of the probability weighting function. Finally, the probability of choosing the risky option for a given subjective value (Eq. 1) is computed using a logistic choice rule

$$p(u) = \frac{1}{1 + e^{(-\tau u)}}, \tag{5}$$

Where τ is an inverse temperature parameter representing the degree of stochasticity in the choice process. All parameters were estimated using a Hierarchical Bayesian approach that uses the aggregated information from the entire population sample to inform and constrain the parameter estimates for each individual (*70*). The hierarchical structure contains two levels of random variation: the trial and participant levels. At the trial level, choices were modelled following a Bernoulli process

$$y_{(p,i)} \sim \mathbf{Bern}(p(u)),$$

with indices (p) for participant and (i) for trial (note that indexes within parenthesis correspond to specifications of the hierarchical level of the Bayesian model). At the participant level, the prospect theory parameters were constrained by group level hyper-parameters. The parameter θ was restricted to be between 0 and 1 and was parameterized using a Beta distribution, which is common practice for parameters limited to values between 0 and 1.

$$\theta_{(p)} \sim \mathbf{Beta}(\mu_\theta \times \kappa_\theta, (1 - \mu_\theta) \times \kappa_\theta),$$

Where hyper-parameters μ_θ and κ_θ represent the mean and the precision of the beta distribution. All other parameters at the participant level were parameterized using normal distributions (with mean = μ and SD = σ) and restricted to positive values where necessary

$$\lambda_{(p)} \sim \mathbf{Normal}(\mu_\lambda, \sigma_\lambda)$$

$$\gamma_{(p)} \sim \mathbf{Normal}(\mu_\gamma, \sigma_\gamma)$$

$$\tau_{(p)} \sim \mathbf{Normal}(\mu_\tau, \sigma_\tau)$$

For latent variables at the highest level of the hierarchy (hyper-group parameters), we assumed flat uninformed priors (i.e., uniform distributions). Posterior inference of the parameters in the hierarchical Bayesian models was performed via the Gibbs sampler using the Markov Chain Monte Carlo (MCMC) technique implemented in JAGS (*70*, *71*). A total of 50,000 samples were drawn from an initial *burn-in* sequence, and subsequently a total of 50,000 new samples were drawn using three chains (each chain was derived based on a different random number generator engine, using a different seed). We applied a *thinning* of 50 to this sample, resulting in a final set of 1,000 samples for each parameter. This thinning assured that the final samples were not auto-correlated for all of the latent variables of interest investigated in the study. We conducted Gelman-Rubin tests for each parameter to confirm convergence of the chains. All latent variables in our Bayesian models had a Gelman-Rubin statistic of less than R<1.05, which suggests that all three chains converged to the target posterior distribution.

3.2. Delegation task decision model description.

The responses in the Delegation task consisted of a set of three possible alternatives: [Defer (d), Risky (r), Safe (s)]. At the trial-wise level, choices were modeled using a conditional regression model. Note that the modeling framework ultimately generates conditional probabilities for each of the three options. For the sake of clarity, we outline the model as a series of computational steps. However, this does not mean that we make assumptions about the temporal order of these computations – they may occur in parallel or serially). The conditional probabilities for each of the possible choices are denoted as follows

$$\begin{aligned} \phi_d &= p(d|u), \text{ see Eq. 7 below} \\ \phi_r &= p(r|u) \times (1 - \phi_d), \text{ see Eq. 8 below} \\ \phi_s &= (1 - \phi_d) \times (1 - \phi_r) \quad . \end{aligned} \tag{6}$$

The probability of selecting the risky (ϕ_{r}) or safe (ϕ_{s}) actions conditional on leading (i.e., 1 – ϕ_{d}) can be computed for a given *u* by using equations 1-5. Therefore, we are left with the task of specifying ϕ_{d} , which defines the probability of deferring given the risky, safe and defer options. We took inspiration from computational models of perceptual categorization (*17*) to estimate ϕ_{d} in choices from our Delegation task.

In both Group and Self trials, participants must decide whether to lead or defer based on the current information, in other words, the difference in subjective values between the risky and safe options. Because the safe option is always zero, this difference is equal to the subjective value of the risky option. Thus, participants need to distinguish (i.e., categorize) between cases in which they prefer to lead (l) or defer (d) given the subjective value of the risky action. In this categorization problem, the participants have to determine whether the subjective value difference is sufficiently close to zero (representing more difficult choices; see main text) such that it is more sensible to defer, or is far enough from zero that it is subjectively preferable to lead.

3.2.1 Lead or Defer (LD) model description. The type of categorization problem outlined above can be modeled by applying a decision (category) boundary criterion κ, to a noisy input measure which in our case corresponds to the u in any given trial. It has been previously demonstrated that an efficient way of solving such problems involves the decision maker comparing two overlapping Gaussian probability distributions (*17*) with a common mean but distinct variances. In our case the behavioral data suggests that the first distribution (representing the internal belief relating to when the individual should lead) is a wider distribution resulting in a relatively high probability of leading for extreme u (i.e. when one option is clearly better than the other). In contrast, the second distribution (representing the internal belief relating to when the individual should defer) is a narrower distribution resulting in a relatively high probability of deferring for u close to 0 (i.e. when the subjective values of the options are similar). In order to make a decision which behavioral option is most appropriate for a given u, an optimal strategy is to compare the log-likelihood ratio of these two probability distributions (*17*). This computation can be expressed in terms of the probability of deferring for a given subjective value difference using the following expression

$$\phi_{\mathrm{d}} = p(\mathrm{d}|u) = \frac{1}{2}\left[\mathrm{erf}\left(\frac{(u+b)+\kappa}{\sigma\sqrt{2}}\right) - \mathrm{erf}\left(\frac{(u+b)-\kappa}{\sigma\sqrt{2}}\right)\right], \qquad (7)$$

where $\mathrm{erf}(\cdot)$ denotes the error function, κ is the optimal category boundary determined by the intersection points of the two probability distributions that are being compared here, and the noise in the representation of u (see equation 8). σ is an estimate of this noise in the representation of u, and b represents a potential bias from zero in the mean of the subjective value distribution. Thus, $p(d|u)$ is determined by an interplay between the distance of a given trial's u from the boundary criterion κ (which determines an indifference point between leading and deferring), and the distance of the mean of the overlapping distributions b. Following previous reports (*17*), we refer to equation 7 as the optimal categorization step in this Delegation model. Note that the probability of leading is simply one minus the probability of deferring, $p(l|u)$= 1- $p(d|u)$.

The optimal category boundary (κ) is set as a function of the participant's prior beliefs and the precision/noise in her representation of the subjective valued difference between the risky and safe options. In our formulation of the model, we estimate the κ parameter directly, but we give the underlying derivation of κ here to make clear how it related to prior beliefs and SVd representation noise.

$$\kappa_1 = 0.5 \log \frac{\sigma^2 + {\sigma_l}^2}{\sigma^2 + {\sigma_d}^2} + \log \frac{p_1}{(1 - p_1)}$$

$$\kappa_2 = \frac{{\sigma_l}^2 - {\sigma_d}^2}{2\,(\sigma^2 + {\sigma_d}^2)\,(\sigma^2 + {\sigma_l}^2)} \tag{8}$$

$$\kappa = \sqrt{\frac{\kappa_1}{\kappa_2}},$$

The σ_D and σ_L parameters represent the SD of the priors over the utilities of deferring and leading, respectively, as a function of the range of SVds in our experiment. Both distributions have a mean equal to zero. The p_1 parameter is the probability that the SVd in a randomly selected trial from the choice set indicates that it's best to defer and because the participants know that the same decision problems are presented in Group and Self trials the final log ($p_1/(1-p_1)$) term cancels out when comparing across conditions. The estimation of only κ and σ directly is beneficial in this case because, while we can infer from the choice patterns that the width of $\sigma_L > \sigma_D$, the exact width of the prior distributions over leading and deferring is unknown. However, please note that κ and σ are identifiable (see **Table S7**).

Having defined ϕ_d, Eq. 6 is fully described by computing ϕ_r which is the probability of choosing the risky option (independent of deferring) already described in the prospect theory modeling section (See Eq. 5).

$$p(r|u) = \frac{1}{1 + e^{(-\tau u)}}, \tag{9}$$

Having specified ϕ_{d}, ϕ_{r} and ϕ_{s}, we assume that the observer makes a categorical decision (defer, risky or safe) following a multinomial distribution.

Thus, we combine the LD and PT models to estimate the parameters for the Delegation task conditional regression (Eq. 6) using a Bayesian hierarchical framework. We fit the model to the data in two ways, placing different levels of constraint on the prospect theory parameters. In our restricted model, we estimated the LD model parameters (i.e., c, b and σ) from the Delegation task choices and fixed the PT model parameters to be equal to those estimated from choices in the separate baseline task for each participant. This method has the advantage of being able to fit the PT parameters to the entire set of risky choices because participants had to make every decision themselves in the baseline task (i.e., they could not defer). However, we also simultaneously estimated LD and PT model-parameters from the Delegation task (full model). Both approaches result in similar findings (see Supplementary 9, **Fig. 4** and **Fig. S6-S7**), but the full model makes no assumptions of stable PT parameters across choice contexts.

In both approaches, at the trial level, choices were parameterized following a multinomial distribution

$$y_{(p,i)} \sim \mathbf{Multinom}(\{\phi_{\mathrm{d}}, \phi_{\mathrm{r}}, \phi_{\mathrm{s}}\}).$$

Moreover, given that in the present study we were interested in comparing Group versus control decisions, we included in our hierarchical structure a simple linear model to capture the effects of Group over control decisions in the following way for the ϕ_{d} parameters

$$c_{(p,i)} = c_{\mathrm{control}(p)} + (\beta_{c(p)} \times \delta_{(p,i)})$$

$$b_{(p,i)} = b_{\mathrm{control}(p)} + (\beta_{b(p)} \times \delta_{(p,i)})$$

$$\sigma_{(p,i)} = \sigma_{\mathrm{control}(p)} + (\beta_{\sigma(p)} \times \delta_{(p,i)}),$$

where $\delta_{(p,i)}$ is a dummy variable indicating whether trial i was a Self trial, $\delta_{(p,i)} = 0$, or a Group trial, $\delta_{(p,i)} = 1$. Thus, the parameter β on each of these expressions indicates for each participant the effect of the Group trials on each of the ϕ_{d} parameters. For the full model, we also included similar expressions for each of the parameters of the prospect theory model. The ϕ_{d} parameters (c , b and σ) and the estimated influence of the group trials (β parameters) were parameterized using normal distributions with group-level hyper-parameters. The model was fit to the data using the same Bayesian estimation procedures described in the Prospect Theory section.

4. Main effects, regressions and correlation statistics.

For robustness to the assumptions of parametric statistical tests, we used the non-parametric Spearman and the Wilcoxon signed rank test approaches for all bi-variate correlations and group comparisons (*72*). For linear regressions, given the potential sensitivity of this analysis to extreme values, we tested for potential outliers via the modified Thompson tau technique [α=0.0001 (*73*, *74*)]. Only one responsibility aversion value exceeded the exclusion threshold (relevant for the regression analysis with responsibility aversion as a dependent variable depicted in Table S1 and the graphical visualization in Fig. 5A). Excluding this highly responsibility averse participant from the analysis of the group mean yielded a result similar to that reported in main text (mean=11%; z=2.63, p=0.009). This original group participant was not assigned to fill in the LBDQ goal oriented leadership questions and thus could not affect any leadership related analysis. We could however further test the robustness of this effect in two ways. First, we excluded this data point from the regression analysis with responsibility aversion as a dependent variable. Second, we report the results of non-parametric (*75*) regressions, including all subjects, which yield similar findings (see **Table S1** legend). The statistical inferences drawn from all tests are based on two-sided p-values and adjusted for multiple comparisons where appropriate. Homoscedasticity was assessed where appropriate via the Bartlett test and no differences in variance were found (p>0.96 uncorrected).

5. Magnetic resonance imaging acquisition and analysis.

5.1 Image acquisition and analysis.

Imaging was performed using a Philips Achieva 3T whole-body scanner (Philips Medical Systems). All images were acquired using an 8-channel Philips sensitivity-encoded (SENSE) head coil 8-channel. Three-dimensional T1-weighted anatomical scans were acquired with high resolution (3D MPRAGE T1 sequence 1 mm^3 voxels). For BOLD scanning, T2*-weighted images were acquired using the following parameters: time until repetition (TR) 2204 ms, Eco time (TE) 30 ms, Flip angle 90º, 37 oblique slices with 0.5 mm gap, -20º from AC PC, 3 × 3 × 3 mm voxel size covering the whole cerebrum.

Statistical Parametric Mapping (SPM 12; Wellcome Trust Centre for Neuroimaging, London, UK, http://www.fil.ion.ucl.ac.uk/spm) standard pipeline was used to pre-process the MRI data. Specifically, after discarding the first 5 volumes to allow for scanner equilibration, images were realigned, unwarped and slice-time corrected (to the middle slice acquisition time). T1-weighted structural images were co-registered with the mean functional image and normalized to the standard T1 MNI template based on the Montreal Neurological Institute (MNI) reference brain, using the segment procedure provided by SPM 12. The functional images were then normalized to a standard EPI template using the same transformation and spatially smoothed with an isotropic 6 mm full width at half maximum (FWHM) Gaussian kernel.

Correction for physiological noise was performed via RETROICOR (*76*) using Fourier expansions of different orders for the estimated phases of cardiac pulsation (3rd order), respiration (4th order) and cardio-respiratory interactions (1st order) (*77*). The corresponding confound regressors, in addition to head movement confound regressors, were created using the Matlab PhysIO Toolbox (open source code available as part of the TAPAS software collection: http://www.translationalneuromodeling.org/tapas/) (*78*). Five functional runs (out of a total of 172) included head motions exceeding 3 mm. Three of these included multiple shifts and were not used in the analysis, resulting in three participants with three functional runs each rather than four. The two additional participants had a highly temporally localized head motion (one spike less than two TRs in length). Thus for these participants, the three TRs before and one TR after this movement were each modeled as a separate regressor and brain activity from this period was not included in the reported analyses. For one subject the scanner terminated a functional run early due to a technical malfunction. For this subject, 264 instead of 280 trials were available.

5.2 Individual level GLMs.

For the fMRI analysis we focus on the critical time period where individuals combine prior knowledge with new evidence to form a decision. For each participant, a time series was created indicating the temporal position of the different trial types in order to compute two general linear models (GLMs) of participants' decisions, including both directly observable variables (i.e., choice type and outcome) and model-derived latent choice variables (i.e., u and p(l|u)). In the primary GLM (GLM-1), choice onsets were divided into regressors for 1) Group trials and 2) Self trials. These regressors were modeled as a boxcar from the time the question was presented until the participant responded. Three parametric regressors containing trial-specific values were added for each of these conditions. 1. RT, 2. *SVd* and 3. p(l|u) For further details on the model-based parametric regressors, *SVd* (subjective value difference, i.e. the output of the PT model in Eq.1) and p(l|u) (the probability of deciding to lead on that trial i.e., the output from our LD model in Eq. 7), see computational modeling section above. Note that these regressors explained mostly independent variance and displayed a weak correlation (rho=0.196). Two additional regressors were created for the period of the feedback in Group and Self trials separately. These regressors were modeled as a 2-second boxcar matching the feedback presentation duration. All regressors were convolved with the canonical hemodynamic response and then entered into a GLM with the BOLD time series in each voxel as the dependent variable.

The second GLM (GLM-2), was very similar to GLM-1, except that the decisions were separated into four regressors (instead of two) as a function of two categorical factors, task condition (Group, Self) and choice type (relying on the group's decision vs deciding alone). The purpose of GLM-2 was to test any categorical differences between trials in which participants ultimately decided to use the group information versus trials in which they didn't. However, we wish to point out that any results for these contrasts do not necessarily reflect post-choice changes in brain activity; in fact, given the timing of the regressor onsets, these differences are more likely to represent prechoice activity that drives choices to make the decision alone (i.e. Lead) or to take advantage of the additional information available at the group level (i.e. Defer) (see GLM-3 below). Once again, the regressors were modeled as a boxcar from the time the question was presented until the participant responded. Two parametric regressors containing trial-specific values were added to each of the four choice-type regressors to account for BOLD signal variance associated with reaction times (RT) and subjective value difference. The regressor for $p(\mathrm{l}|u)$ was omitted due to lack of variance within the separate Lead and Defer trials. Four additional regressors were created for the period of the feedback corresponding to the aforementioned regressors for the time of the decision itself. These regressors were modeled as a 2 second boxcar matching the feedback presentation duration. All regressors were convolved with the canonical hemodynamic response and then entered into a GLM with the BOLD time series in each voxel as the dependent variable.

Lastly, we estimated a third GLM (GLM-3) that was identical to GLM-2 except that we added a parametric regressor indicating the informational advantage of deferring to the group. The informational advantage increases as function of the amount of covered bins (ranging from 1-9) (see **Fig. S1**). We entered the number of covered bins per trial as a trial-wise parametric regressor in GLM-3 for all four trial types (Group, Self, Lead, and Defer).

All parametric regressors in each GLMs were sequentially orthogonalized such that any shared variance was assigned to the preceding regressors rather than the regressor of interest (consequently, factors controlled for in each analysis were always positioned before the regressor of interest, see SPM12; Wellcome Trust Centre for Neuroimaging, London; http://www.fil.ion.ucl.ac.uk/spm). Reaction time was always added as the first regressor to remove any potential confounds related to this factor in the interpretation of the results for *SVd* and $p(\mathrm{l})$.

5.3 Group level analyses.

Single-subject contrasts were computed following the GLM analysis and used in standard random-effects group-level analyses (t-tests). The individual-specific value of responsibly aversion was also added as a covariate in the random effects analyses. All results are whole-brain corrected for multiple comparisons (see **Tables S5-S7** for details).

5.4. Dynamic causal modeling (DCM) network analysis.

We used a non-liner stochastic DCM approach (*28*, *79*) to estimate the functional coupling within a four-region network in which each region was linked to a separate task or decision component (**Table S3**). The primary purpose of this DCM analysis was to examine associations between interregional functional connectivity and behavior (i.e., responsibility aversion and leadership). Thus, our aim here differed from the more common utilization of DCM in order to compare different possible models of brain connectivity. Therefore, while many steps of our DCM analyses correspond to those presented in the previous literature (*80*, *81*), there are key differences due to our specific aim. This approach provides several key advantages for exploring inter-individual differences because it allowed us to test which aspects in our minimalistic network correlate with the behaviors of interest while avoiding potential bias in model selection and differences in optimal models across individuals.

As a first step, we extracted four activation time courses from functional masks in medial Prefrontal Cortex (mPFC), Anterior Insula (aIns), superior/middle Temporal Gyrus (TG), and Temporal-parietal junction (TPJ) (**Table S3**). The time courses were extracted from a 5mm sphere centered on each individual participant's peak for the relevant contrast (*SVd*, $p(\mathrm{l}|u)$, Group>Self condition and Defer>Lead trials, respectively). The individual peak was selected from within a 10 mm sphere centered on the group-level contrast peak (identified using both parametric and non-parametric methods see **Table S3** and legend). Note that we combined information from the current univariate statistical results with previous evidence from other studies when deciding which regions to include in our DCM. Thus, the most significant peak in our current results was not automatically selected for inclusion in the DCM, but instead we made principled decisions in favor of using certain regions and not others based our knowledge of brain structure and function (for full activation list see in **Tables S3-S5**). Specifically, we opted not to use the occipital areas identified in the Group > Self contrast because we believed that they may be more related to visual attention processes than responsibility per-se (*82*) and instead included the MTG/STG region that has been found to be related to the ability to distinguish between aspects of the self and of others (*83, 84*). Similarly, given prior evidence for the neural representation of SVd signals (*29, 30, 85, 86*), we selected the mPFC region for the SVd contrast. For simplicity, we selected the left hemisphere peak when bilateral activity was identified (aIns). In the current study, we restricted ourselves to a single set of regions for our DCM and subsequent machine learning analyses. The construction and comparison of alternative models with other regions is an interesting avenue for future research, but such exploratory analyses were not our goal in the current report.

Second, we specified a DCM including the relevant inputs into each region (corresponding to the specific contrast used to identify the region and all the possible connections between regions). We allowed the connections between the regions to be modulated according to the different conditions in the task (Group, Self) and ultimate decision outcomes (Lead, Defer). We also allowed for non-linear effects of activity in the TPJ region where activity was greater when participants utilized the groups perspective and the TG region that was more active in Group compared to Self trials, on the coupling between the mPFC (reflecting *SVd*) and aIns (reflecting $p(l)$). These non-linear effects were included to potentially parallel findings from our computational model that suggested differences in the relationship between *SVd* and lead decisions (i.e. the *SVd* magnitude participants required before choosing to lead) between Group and Self conditions. This DCM was estimated separately for each participant and the parameters used in the prediction analysis described below.

As aforementioned, because our goal was to use the DCM parameters to predict the responsibility aversion and leadership scores of independent test participants, we opted not to perform model comparison or selection between DCMs to avoid any inadvertent bias or over-fitting to the training samples that would reduce generalizability. However, for completeness, we additionally provide the results of a model comparison platform using the DCM optimization algorithm implemented in SPM12 (*87*, *88*). This procedure uses a Bayesian model comparison method to provide group level parameter estimates after model selection on the individual subject level (**Fig. S8**). The DCM model performed well in explaining the data and accounted for a high percentage (group-level average = 42.3% ±0.48) of the variance in the total (task and non-task related) activity in the four regions of interest.

6. Predicting responsibility aversion and leadership scores using DCM parameters.

We used an elastic net regression in combination with a leave-one-out cross-validation approach (*89–92*) to examine the relationship between the neural network parameters estimated with DCM and the participants' responsibility aversion levels and leadership scores. The elastic net regression procedure optimizes a standard criterion for a within-sample-fit that is subject to a penalty which increases monotonically as the sum of the coefficient vector increases in absolute magnitude. This procedure thus penalizes for the use of many non-zero coefficients (i.e., it selects only the most relevant parameters) or for assigning high-magnitude coefficients (i.e., it shrinks parameter coefficients towards zero). The shrinkage and selection properties of the elastic net serve to reduce the likelihood of over-fitting to the training data set. For the elastic net analysis we used an alpha value of 0.3 following optimization performed in previous work from our lab (*93*). The leave-one-out cross-validation procedure first uses the data from n-1 participants to estimate model parameters. Subsequently these parameters are employed to predict a score for the independent participant that was not included in the n-1 analysis. An n-fold replication of this procedure produces n predicted (i.e., one for each participant) values that can be compared to the real values of each participant to determine accuracy. Here we quantified accuracy in terms of both classifying values as high or low, relative to the training set median, and the correlation between predicted and true values for leadership and responsibility aversion. Control analyses indicated that there was no unintended bias in these classification procedures because randomly reshuffling the training set labels yielded chance classification rates and no correlation between the predicted and true values of responsibility aversion or leadership scores (see Supplementary 10.3).

The elastic net regressions included all DCM parameters [all inputs, intrinsic connections, and relative modulations of interest on these connections (Group-Self), and nonlinear effects] to relate our estimates of neural coupling to leadership scores. To test the additional explanatory power of our DCM parameters, we compared between models using only behavior or behavior and DCM parameters to explain the composite leadership scores. The first model included all parameters listed in **Table S1** as regressors. The second model was identical but additionally included all DCM parameters mentioned above as regressors. Model comparisons using both the AIC and BIC (*94*, *95*) penalties for model complexity indicate the superiority of the model including behavior and DCM parameters (AIC and BIC difference 194.5 and 132.8 respectively). We further quantified the predictive power of our neural connectivity model parameters using the leave-one-out cross-validation approach (*89*).

II. Supplementary Results

1. Higher leadership scores were not associated with sensitivity to the informational advantage of deferring.

Each participant individually answered the same questions (identical probabilities, cover size and potential gains and losses) in the baseline choice task in Stage 1. However, the location of the covered portion of the probability pie varied for each member of the group (**Fig. S1A**). By varying the position of the gray cover for each individual in the group, we ensured that each individual saw the same amount but not the same content of information. Specifically, the cover position for each group member was pseudo-randomly assigned to ensure that on every trial no two participants saw the exact same information. Thus, the total amount of information across all group members was larger than that available to each individual alone. Given this additional information, there is a corresponding increase in the expected accuracy of the average answer of the group compared to that of each individual [**Fig. S1B**, this task feature is related to the "wisdom of the crowds" concept (*96*, *97*)]. The size of this informational advantage increases as each group member's level of private information decreases (i.e., as ambiguity increases). In other words, emulating real-life situations, when an individual has a large amount of information on her own, she stands to benefit less from an additional perspective. At the beginning of Stage 2, this design feature and its implications were fully explained to all participants. Critically, the informational advantage was identical for the matching Group and Self trials. Indeed, although the informational incentive did prompt participants to defer more as the ambiguity increased (**Fig. S1C**), this effect did not differ between the Group and Self conditions for either experimental group (a model regressing the percentage of deferring on the size of the gray cover in each experimental condition did not reveal a difference in slopes between these two conditions; interaction term F(7)=0.2; *p*=0.98).

One reason that we included the informational advantage of deferring was to offset individuals' preference for control, or the desire to retain decision rights for themselves (*5*). The fact that participants prefer to retain the right to make decisions themselves is demonstrated by the fact that deferral rates are well below 100% (which is the profit maximizing strategy) (mean deferral= 37.4% and 41.5%; non-parametric sign test vs. a random-choice null-hypothesis, sign=13, p=6*10^{-10}, and sign=22, p=2*10^{-4}, for the Self and Group trials respectively).

Our procedure creates an ecologically valid trade-off between preferences for control and gaining additional information, an important feature of realistic decision environments (*8*, *98*) that also reflects realistic individual preferences for leading. This is because individual group members' opinions matter and can potentially lead to a better decision. The informational advantage assures that individuals have an objective reason for deferring and will do so on both the Self and Group conditions, thus ensuring sufficient variance for the analyses. Varying the level of informational advantage in a parametric fashion allows us to directly test how much the individuals are willing to "pay" by forgoing a resource (in this case information) in order to maintain their decision rights in the Stage 2 delegation choice task. Moreover, this design allowed us to test whether the informational advantage factor interacted with responsibility aversion in leadership, thereby providing a test of the hypothesis that better leaders defer more when doing so is advantageous for achieving the objective.

We can test this hypothesis by examining whether individuals who score high on leadership measures defer more when they lack extensive private information, thereby increasing the chance of success due to the more extensive information available to the group. This would result in a sharper slope in the relationship between deferring and ambiguity in better leaders. However, we did not find that to be the case in either experimental group. For example, regressing leadership scores on the slopes in the Self and Group conditions, as well as their interaction, did not reveal any significant effects (across leadership measures for all participants; all p>0.5; not corrected for multiple comparisons). Moreover, a direct comparison between the slopes in the Self and Group conditions failed to find a significant difference (Wilcoxon signed rank test; z=-0.1, p=0.92) suggesting that differences in the slopes did not explain the behavioral differences between these conditions. In summary, although the informational advantage of deferring is an important factor in participants' decisions, its effects on choice did not change as a function of responsibility and were unrelated to leadership characteristics.

2. Choice consistency across decisions is not associated with leadership scores.

Our measure of responsibility aversion is directly related to how consistently an individual makes choices in the Self and Group conditions of our task. Therefore, an alternative explanation is that we may be indexing choice consistency rather than a change in decision policies in the face of responsibility. However, choice consistency would result in similar response patterns when answering the same questions in the baseline test and the Self condition (which does not involve responsibility). Although in the latter case individuals have the additional option to defer, on a large number of trials (88.3±2.3 trials on average) they choose either the risky or safe options, allowing us to compare the two sets of choices. The correspondence between the answers in these two sets (as measured by the Phi Coefficient for association between binary vectors) can be taken as a proxy for general choice consistency. As expected, the average choice consistency was high but varied substantially across individuals (mean ϕ= 0.68, range 0.21-0.95).

There was no association between choice consistency (ϕ) and leadership scores score in either group (for brevity across all participants *rho*=0.06 and 0.01 for LBDQ and field measure respectively, both p>0.65 uncorrected). In addition, no relationship with choice consistency was observed for the Composite leadership score in the fMRI replication group (*rho*=-0.19 p=0.25 uncorrected). These results suggest that simple choice consistency is not the relevant feature of leadership decision making captured by our responsibility aversion measure.

3. Out-of-sample prediction of leadership scores.

We computed an out-of-sample forecast of the leadership scores in the fMRI replication group using the parameter estimates from the original group. Specifically, we estimated an elastic-net regression using the variables listed in **Table S1** to explain individuals' leadership scores in the original group's data. We then used the estimated coefficients to create a predicted leadership score for every individual in the fMRI replication group based on their individual values on the measures listed in **Table S1**. For one fMRI replication-group participant who did not complete the in-group affiliation score, the prediction was obtained by refitting a second model to the original group data using all factors except the in-group affiliation score and then predicting the leadership score based on this reduced model. Next, we compared the predicted leadership scores to the actual scores obtained from the leadership questionnaires in the fMRI replication group and found that the predicted leadership scores were significantly correlated with the scores obtained from the LBDQ (Spearman *rho*= 0.44 p= 0.004). We used the LBDQ rather than the composite score of the fMRI replication group in this prediction analysis because only the LBDQ was available from the Original sample and we wanted a direct, 1:1 comparison.

Lastly, to test if the predictive accuracy of the model stems from responsibility aversion or some combination of the other regression variables, we conducted the same analysis using a reduced version of the model without the responsibility aversion regressor. The resulting correlation between the predicted and actual scores was substantially lower (*rho*=0.25, *p*=0.12), suggesting that responsibility aversion preferences were a key component of the model. A direct comparison between the full and nested models revealed that the model with the responsibility aversion parameter explains significantly more variance in the leadership scores (F-test for nested models F(1,27)=11.2 *p*=0.003).

4. Responsibility aversion did not significantly correlate with traditional psychological traits assessed via the Big 5 inventory.

Our goal in this work was to examine individual differences in the decision process underlying leadership choices. When discussing our results with leadership researchers they raised the question of whether or not responsibility aversion is related to general psychological traits (e.g. openness or agreeableness). Therefore, we re-contacted our participants and asked them to fill out The Big 5 personality questionnaire (*99*) for an additional payment of 20 CHF. We received responses from 49 participants across both samples and used this data to test the ex-post questions about associations between general psychological traits and responsibility aversion. None of the big 5 personality scales were even marginally correlated with responsibility aversion in bi-variate tests (extraversion, rho=-0.12, p=0.5; agreeableness, rho=0.03, p=0.88; conscientiousness, rho=-0.04, p=0.8; neuroticism, rho=0.22, p=0.21; openness, rho=-0.08, p=0.64). We also entered all big5 scores simultaneously into a multiple regression model with responsibility aversion as the dependent variable. None of the personality scales was significantly associated with responsibility aversion in this model either (all p>0.18, uncorrected). Thus, responsibility aversion appears to be a unique choice preference that is not captured by other, more traditional choice preferences or standard personality traits.

5. Preferences over regret, guilt and accountability or blame do not explain responsibility aversion; additional analyses and an additional control experiment.

Previous research has found that individuals tend to avoid choices they believe will lead to regret or guilt in the future (*100–103*). Regret is experienced *"when it turns out, in retrospect, that you should have chosen something different"* (*104*). In our task, however, participants always face the same risky prospect in the matched Self and Group trials and, therefore, regret cannot affect choices differently in the two trial types. Thus, regret for not choosing the correct response cannot be a driving factor behind the responsibility aversion we observe.

Guilt can be *"formally operationalized as failing to live up to another's expectations"* (*101*, *105*) and because losses typically loom larger than gains, feelings of guilt may be particularly salient if one's decision imposes losses and potential harm on others (*106*). Therefore, if avoiding guilt is an important driving factor for responsibility aversion, we would expect that participants defer most often when potential losses were greatest (i.e., when the potential harm was highest). However, our results indicate that responsibility aversion did not increase for events where potential losses were larger than potential gains. In fact, we find the opposite pattern because the mean deferral rate in the Group trials was 45.3±1.8% for |loss| < |gain| but only 33.4±1.9% for |loss| > |gain| (z=5.6, p=3*10^{-8}).

In addition, we conducted a separate control experiment to test whether reducing the personal accountability eliminates the responsibility aversion effect. To that end, we collected behavioral data from an additional 32 participants (24 females, mean age 22.7±0.74) who underwent the same protocol as described in the Methods section with one key difference. Specifically, accountability was minimized by providing no feedback regarding the outcome of participants' own choice or the choice of their group after each trial, nor at the completion of the experiment. Participants received only a total payment at the end of the experiment and were not informed which portion of the money was earned from their own versus their fellow group members' task performance. Recall, that each individual is paid out for her role as the potential leader as well as the role of group member when the other three group members have the leadership role. Therefore, neither the participant nor any of her group members could know how well she performed for herself or the group. Using this modified protocol, we again find a large (28.9%) and significant (*z*=3.5; *p*=0.0005) increase in deferral behavior in the Group vs. the Self condition. Taken together, these results suggest that neither guilt nor blame is likely to be the main driving factor of responsibility aversion.

6. Participants do not defer to align the choice strategy with the preferences of other group members.

Another potential contributor to responsibility aversion is a preference for giving the other group members the ability to determine their outcome according to their own preferences. We term this a democratic preference, but do not find convincing evidence supporting this as the main motivation for responsibility aversion. Under this assumption, we should observe that deferral choices are correlated with the participant's perception of the other group members' expectations. In other words, when one's risk preferences are very different than the group's, there is a higher chance than if they make the choice themselves, they will disappoint the group and thus they should defer. However, neither the actual (Table 1) nor the perceived expectations of the other group members were correlated with responsibility aversion. In the fMRI replication group we elicited ratings from each participant regarding how similar they are to their group members in terms of risk preferences (*107*, *108*) at the end of the experiment. Individuals were asked to rate their relative position in terms of risk taking preferences vs. their group on a 1-5 scale (ranging from "On average the group prefers to take more risks than I do" to "On average I prefer to take more risks than the group"). The absolute difference in the risk preferences between the individual and the group did not correlate with responsibility aversion (*rho*=0.11 *p*=0.47 uncorrected), suggesting that perceived differences in risk preferences between the individual and the group are not driving the responsibility aversion effect.

In addition, if democratic concerns were driving decisions, this would not be limited to defer choices and would be evident when deciding to lead as well. For example, when making a leadership choice between a risky and safe option with direct effect on others, the others' preferences should be taken into account. Thus, as an additional test we entered the others' average risk and loss preferences into a logistic regression with the dependent variable as the selection of the risky or safe options from lead choices (on relevant trials in which individuals can affect others, i.e., Group trials). No significant effect was found for these factors (all *p*>0.5 uncorrected).

7. Response times are similar in Group and Self trials and do not correlate with leadership scores.

We tested whether response times differed as a function of treatment condition or leadership scores. Response times did not significantly correlate with leadership scores overall or across the different conditions (max rho=-0.15, p=0.25 uncorrected). Furthermore, there were no significant differences in mean response times across conditions. A two-way ANOVA with Condition (Group/Self) and Response (Lead/Defer) as factors, reveled a main effect of response (F(1,316)=10.8, p=0.001) but not of Condition (F(1,316)=0.3, p=0.59) or the interaction of Condition and Response (F(1,316)=0.17, p=0.68). The average RT was faster on Lead vs. Defer trials (5.4 vs 6.2 seconds respectively) but was very similar between Group and Self trials (5.7 vs 5.6 seconds). Lastly, a two-way ANOVA with Condition (Group/Self) and absolute subjective value (divided into 5 bins representing the bins in **Fig. 3** collapsed across sign) as factors, did not show a significant interaction (F(4,10)=0.28, p=0.88) either. Nevertheless, we account for RT in all fMRI analyses by including it as the first parametric modulator in both the GLMs (see Methods section 5.

8. Participants decide based on the subjective value of individual payoffs in both the Self and Group conditions.

In order to maintain an identical monetary incentive in the Group and Self trials, and in line with procedures established in the literature (*14*, *15*, *37*, *38*, *103*, *109–111*), the potential outcome for the individual was equated in both conditions. Otherwise, if the potential reward for the whole group in Group trials would have been equated with the individual reward in the Self trials, individuals could have made only a fourth as much for themselves alone on Group vs. Self trials, creating strong differences in incentives. Consequently, although unlikely, it is possible that participants may have perceived their personal utility on Group trials as the combination of their own and part of the others' outcomes (e.g., the individual outcome x 4) in Group trials. Our computational modeling results indicate that this was not the case.

In the full version of our Delegation model (see Methods), the subjective value differences for each choice are estimated directly on the basis of the data from the delegation task. If individuals viewed the utility in Group trials as the reward for themselves and others in the group, then the prospect theory parameters or the stochasticity parameter τ (which is a direct multiplier of the subjective value) would be different in Group versus Self trials. This was not the case (see **Fig. 4C**, **Fig. S6** and Methods section 3).

9. Additional Modeling results.

9.1 Full model of Delegation task choices.

This is the primary version of the model because it allows us to determine if the Delegation task conditions influence any subset of the parameters. This includes the PT model parameters that capture the influence of conventional choice preferences on the estimation of subjective values. In the full model, we allowed the PT parameters describing choices under uncertainty to vary across the Group and Self conditions instead of fixing them to the values estimated from the separate baseline task choices. Thus, all parameters were estimated directly from Stage 2 choice data. As depicted in **Fig. 4** and **Fig. S6**, the model predictions show an excellent fit with the actual data. Using this approach, we find that the only factor to significantly differ between the Group and Self conditions is the deferral threshold. **Fig. 4C and Fig. S6C** depict the change from the Self to Group conditions. The PT values (depicted as means ± s.d.) for the Self condition along with the change in the Group condition (depicted in brackets) were as follows: original group: γ=1.1±0.43 (0.01), λ=1.43±0.31 (-0.01), θ=0.43±0.02 (-0.00) and τ=0.49±0.07 (0.04); fMRI replication group: γ =0.93±0.46 (0.02), λ=1.72 ±0.5 (0.05), θ=0.48±0.02 (0.00) and τ=0.38±0.05 (0.01); for the risk, loss, ambiguity and stochasticity parameters respectively. These results show that the PT model parameters did not change in the Group compared to Self condition, suggesting that responsibility aversion is independent of valuation mechanisms as captured by PT.

9.2 Restricted model of Delegation task choices.

In a secondary, restricted version of our primary Delegation model, we estimated the Prospect Theory parameters from the independent *baseline task* data and only fit the condition-specific deferral threshold and bias parameters to the Delegation task choices. Recall that participants faced the same risky prospects in the baseline and Delegation tasks. However, the baseline task choices provide a wider range of decisions from which to estimate the Prospect Theory (PT) parameters because participants always had to choose the risky or safe action and could not defer to the group, resulting in more detailed information about participants' risk, loss and ambiguity preferences. The resulting parameters as well as the stochasticity in the choice process were all similar to those reported in previous studies of decision making (*18*, *112*). For the original group, the averages over the individual preferences parameters (±s.d) were as follows: γ=0.77±0.25, λ=1.43±0.08, θ =0.43±0.02 and τ=0.32±0.03. For the fMRI replication group the averages (± s.d) over individuals' preference parameter were: γ=1.04±0.1, λ=1.76±0.16, θ=0.42±0.02 and τ=0.22±0.02. As depicted in **Fig. S5** and **Fig. S7** this PT model closely captures the actual choice data from the baseline and Delegation tasks. Using the restricted model we again find that the only parameter to differ between the conditions of interest in either group was the criterion [mean(±SD)=1.2(±0.35) and 1.1(±0.27) with posterior probabilities of a difference between the conditions = 0.99 and 0.99 for the original and fMRI replication groups, respectively].

9.3 Replacing optimal categorization with a conventional logistic choice rule in the Delegation Model.

We also tested two alternative models for computing the probability of defering ϕ_{d} and compared these to the optimal categorization model (see Equation 7 above). In the alternative models, we assumed that ϕ_{d} is computed via a conventional logistic choice rule

$$\frac{1}{1+e^{(-z(u))}},$$

Where *z*(*u*) can take two alternative forms as a function of the subjective value difference (*SVd*). In alternative model 1, we considered a simple linear relationship:

$$z(u) = \beta_0 + \beta_1 |u|$$

Similar to the optimal categorization model, in alternative model 1 the probability of deferring will be higher for values close to 0 (i.e. $|u| \to 0$) with $z(u)$ having a negative slope (i.e. $\beta_1 < 0$).

Alternative model 2, is similar to alternative 1 except that we modelled $z(u)$ using an exponential decay function of the following form

$$z(u) = \beta_0 + \beta_1 e^{-\lambda|u|}$$

As described for the optimal categorization model, we also included in the hierarchical structure of the alternative models a simple linear model to capture the effects of responsibility over control decisions. The quality of the model fits was computed by combining leave-one-out and Pareto-smoothed importance sampling, an approach which has several advantages over alternatives such as DIC (*113*). The optimal categorization model substantially outperformed both alternative models. These results demonstrate that the optimal categorization model explains the leadership decision process better than alternative formulations that combine estimates of the *SVd* with a conventional logit choice rule.

Model fit value	Original Group	Replication Group
Optimal Categorization	13073	15343
LOGIT 1	14919	17464
LOGIT 2	14938	17473

10. Supplementary Imaging results:

10.1 Univariate fMRI analysis protocol and results.

The fMRI replication group performed the second stage delegation choice task while being scanned with Functional Magnetic Resonance Imaging (fMRI). In a univariate analysis of these fMRI data, we identified regions where activity differed as a function of task condition (Group, Self), or choice type (relying on the group’s decision vs deciding alone), as well as regions where activity correlated with the model-derived latent variables *SVd* and *p*(l|u). Thus, we had four contrasts of interest corresponding to the four primary components of each trial. The results for each of the contrasts are listed in **Tables S3-S5**.

To create our minimal neural network and conduct the DCM analysis of inter-regional coupling, we identified four regions or nodes that were associated with each of the four trial components listed above. According to the DCM guidelines (*80*) these regions were chosen based on the statistical results of our GLMs and prior knowledge of neurobiological structure and function derived from previous work (see below and **Table S3** for additional details). Although several brain regions survived whole-brain multiple comparison corrections for each contrast, we used a minimal network by including only a single region from each contrast in our DCM analysis. Moreover, to avoid statistical inference issues related to exploring multiple model spaces in our leave-one-out cross validation predictions, we used only one set of four regions. Therefore, we cannot exclude the possibility that additional regions may also constitute important and/or more predictive nodes in the network. We make no assertion or claim regarding the set of four regions we included being the best for explaining or predicting leadership relevant behaviors. Our objective here is not to characterize the definitive network for decisions in our Delegation task, but rather to demonstrate that a reduced brain model, can be useful in better understanding the process of leadership decisions and in predicting an individual's responsibility aversion and leadership behavior.

The regions we selected correspond to those found to perform similar functions in previous studies. First, subjective value was correlated with activity in medial prefrontal cortex voxels located within the broad mPFC cluster found to be associated with similar computations in a wide range of choice tasks (*29*, *30*). Second, the contrast of the Group and Self conditions activated a region on the border of the Middle and Superior Temporal Gyri (TG) that has been associated with the ability to distinguish between aspects of the self and others (*83*, *84*). Third, relying on the group's decision (i.e. deferring) was most strongly associated with activity in the Temporal Parietal Junction (TPJ), a region classically related to Theory of Mind (i.e., perspective taking) (*114*). Finally, the key estimate of the probability of leading generated from our combined PT and optimal categorization model correlated with activity in the Anterior Insula (aIns), a well-established integration hub proposed to combine internal and external states, together with estimated option values, to inform choice (*60*, *61*, *115–119*).

After selecting our four regions and estimating the functional coupling between them with DCM, we tested if any parameters (i.e. inter-regional connectivity changes or local activity responses) of the network were associated with inter-individual differences in responsibility aversion. Specifically, we examined the relationship between the responsibility aversion measure derived from our model, and all connectivity changes (Group minus Self trials) and direct, task-related effects in each region using a robust regression (FWE<0.05 Bonferroni-Holm corrected for multiple comparisons, **Table S6** and **Fig. 5**).

10.2 Temporal Parietal Junction (TPJ) activity correlates with the informational advantage of deferring to the group.

Higher activity in the TPJ when utilizing the informational advantage (i.e. deferring) could potentially reflect Theory Of Mind processes (*114*) such as the estimation of the possible advantage of acceding to the perspective of others. To further elucidate the function of the TPJ in our task we performed an analysis testing if activity in the TPJ correlated with the degree of informational advantage for deferring. Using GLM-3, we tested whether activity in the TPJ region identified by the contrast of relying on the group's decision vs. deciding alone and used in the DCM analysis, parametrically increased as a function of the informational advantage regardless of whether the participant decided to take advantage of the group level perspective (i.e. on both lead and defer trials). We found that activity in this ROI (averaged across all voxels) was significantly correlated with the level of potential informational advantage on trials that ultimately resulted in both lead and defer decisions. The median beta value averaged across all participants in lead trials was 0.05 ±0.02 (z=4.2, p=0.00003), and in defer trials the mean beta was 0.07 ±0.04 (z=2.9, p=0.004). Given previous evidence concerning the function of the TPJ (*114*) our results suggest that activity in the TPJ may be tracking the informational advantage of taking the perspective of others in our task, and further increase when deciding to rely on others' perspectives to make the choice (i.e. defer).

10.3 Control analysis for the classification of responsibility aversion and leadership scores based on the neural data.

As indicated in the main text, a leave-one-out elastic net classification procedure using the neural network parameters yielded accurate out-of-sample predictions for both the responsibility aversion and leadership scores. As a robustness check we ran the same analyses after randomly shuffling the labels of the dependent variables across the participants (results averaged over 1000 reshuffling simulations). Any better than chance classification in this analysis would suggest a bias not related to our effect of interest, may be present in the classification procedures. This test of the naïve classifier resulted in classification at chance levels (0.492% and 0.499% for the responsibility aversion and leadership scores respectively) and near zero correlations between the predicted and observed scores (*rho*=0.022 and *rho*=-0.017 for the responsibility aversion and leadership scores respectively).

11. The proportion of defer choices does not significantly increase over the experimental time course.

In order to test whether deferring behavior changed as the experiment progressed, we divided the trials into 5 bins according to their temporal order. We then entered the percentage of deferring into a two-way ANOVA with Condition (Group/Self) and Time bin as factors over all participants. There was a significant main effect of Condition (F(1)=11.2, *p*=0.001) but not of Time (F(4)=1.4 *p*=0.23) nor of the interaction of Condition and Time (F(4)=0.1 *p*=0.98). These results do not support a shift in the behavior of interest over time in our task.

12. Reducing responsibility by seeking uncertainty?

A study on decision making under ambiguity suggested that when selecting between two possible negative alternatives, individuals prefer the uncertain option, essentially shifting the blame for the potential negative outcome to chance (*103*). Participants in our task never faced choices between two losses. However, to investigate if uncertainly-seeking interacts with responsibility aversion we tested whether, conditional on leading, individuals prefer the uncertain (risky) choice to a larger degree in the Group versus Self conditions. We do not find this to be the case. In fact, although not significant, individuals tended to choose the risky option to a slightly lesser extent in the Group trials (*z*=-0.76 *p*=0.45).

III. Supplementary Figures.

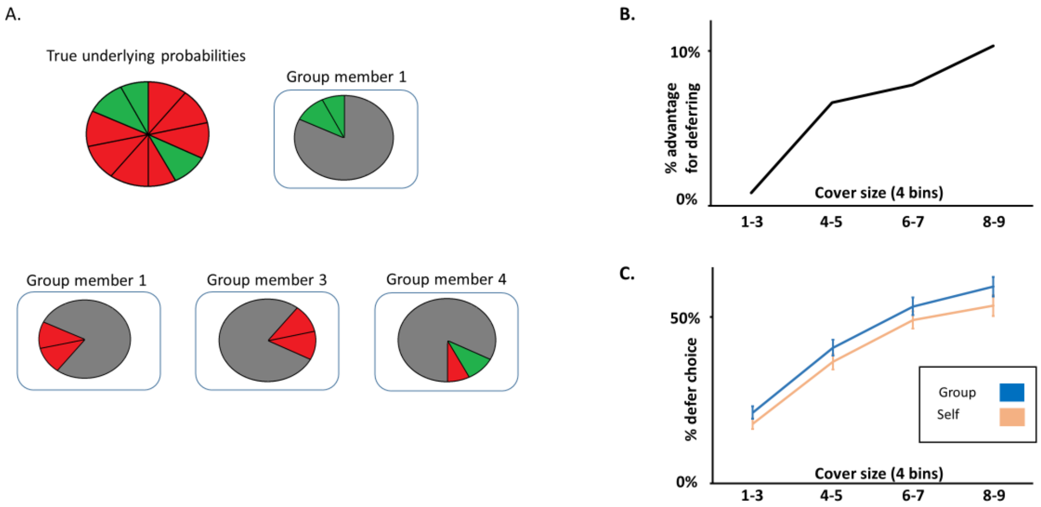


Figure S1. The informational advantage available by deferring to the group consensus. **A**. graphical example of potential observed probabilities seen by each individual in the group as well as the true underlying probability pie, which was not displayed to any participant. **B**. The theoretical informational advantage for deferring to the group vs. retaining control across increasing levels of ambiguity. This advantage is defined as the percent increase in choosing the option most likely to produce a better outcome, provided each individual chooses the option with a higher probability to succeed according to the information available to them. In other words, this measure captures the increased probability of ending up with a gamble when winning is most likely, and the safe outcome of zero when losing is more likely, following deferring vs. acting alone. Emulating real life situations, our task was constructed such that when the individual had a large amount of information herself, she could benefit less from additional information from others. **C**. Participants' deferral behavior. The figure shows the correspondence between the participants' actual behavior and the underlying informational advantage [correlation between average deferral rate and number of obscured bins (1-9), *rho*= -0.97, *p*=8*10^{-5}]. The fact that participants prefer to retain the right to make decisions themselves is demonstrated by the fact that deferral rates are well below 100% (which is the profit maximizing strategy) and decrease when retaining control is not too costly in terms of the informational advantage. Our results suggest that a higher sensitivity to the informational advantage (i.e. sharper slopes) is not reliably associated with responsibility aversion or real-life leadership (see also supplementary results 1). In panels **B** and **C**, trials are divided into four bins based on the level of ambiguity (1-3, 4-5, 6-7 and 8-9 gray slices). Error bars represent s.e.m. across participants. For additional information see supplementary section 1.

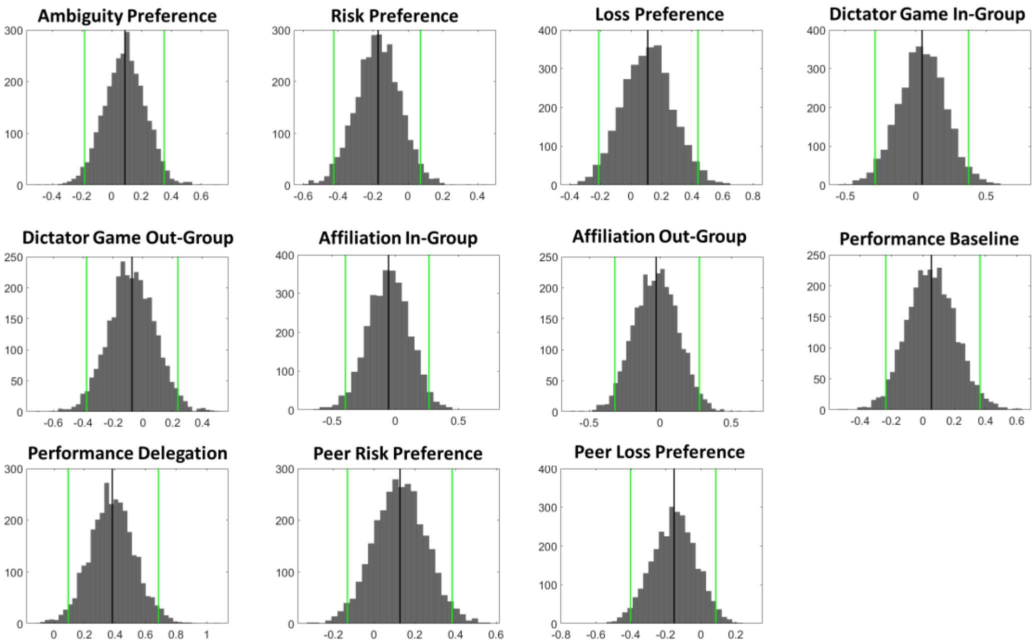


Figure S2. Bayesian posterior distribution for baseline preference measures. Recall that the baseline preference measures were insignificant explanatory variables for the leadership scores in the original and in the replication sample (Tables S1A and S1B). It remains possible however that despite being weaker correlates of leadership than responsibility aversion in our data, given a larger number of observations such preferences may also play a role. To test this we computed the Bayesian posterior distribution (under the assumption of a uniform prior) of the coefficient from a linear regression explaining leadership scores as a function of baseline preferences using all participants in both the original and fMRI replication groups (i.e. combining the data from Tables S1A and S1B). These posterior plots show the relative levels of evidence, across all 81 participants, for each regression coefficient, with zero representing no effect. Green lines indicate the 95% credible intervals and the black line indicates the mean of each distribution. “Performance Baseline” and “Performance Delegation” are total accumulated earnings across all trials. “Peer Risk Preference” and Peer Loss Preference” are the average risk and loss parameters of the other group members.

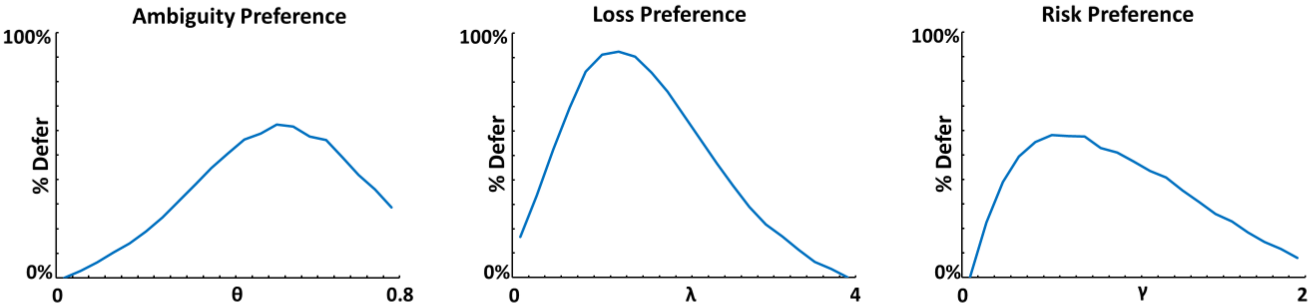


Figure S3. Simulations of alternative mechanisms for responsibility aversion. Graphs of simulations showing that changes in preferences over risk, loss, or ambiguity could theoretically result in the increased level of deferring observed in our task. These plots show the deferral rate as a function of different risk, loss or ambiguity preferences. The deferral rates were computed by averaging over twenty simulated agents faced with the set of ambiguous choices used in our task (note that this set has an overall positive expected value for an agent who is neutral in regard to ambiguity, loss, risk). In each plot, the x-axis shows a range of values for the given parameter and the corresponding deferral rate is given on the y-axis. The simulations and resulting plots varied only one parameter at a time and held all other parameters fixed at the mean value of the group-level parameter estimated from our empirical data in the Self condition.

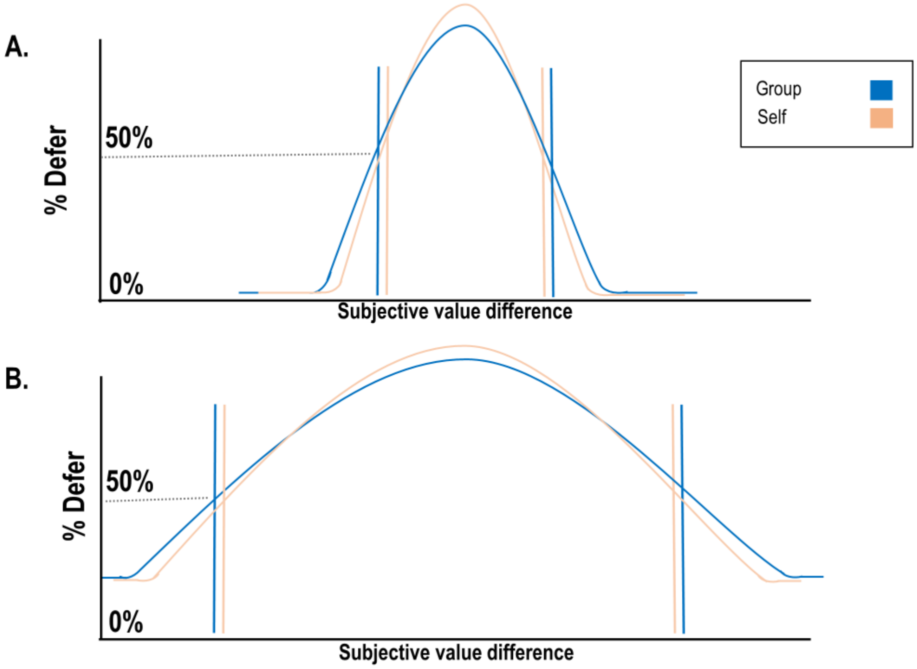


Figure S4. Example representation of autocratic and democratic leaders. These schematic choice curves depict two individuals who make decisions and take responsibility at very different rates, but have equal levels of responsibility aversion. The vertical lines indicate the decision criterion in each choice condition above which the subject will be more likely to lead then defer. The hypothetical individual in panel **A** defers only in a very narrow range of trials with a subjective value difference near zero. This represents an autocratic leader who prefers to lead in most cases. Panel **B**. represents a more democratic leader who prefers to defer to the group vote in most cases and only makes the choice independently when the best action is very clear. Importantly, the responsibility aversion levels in both individuals are equal despite the distinct differences in choice patterns and leadership style.

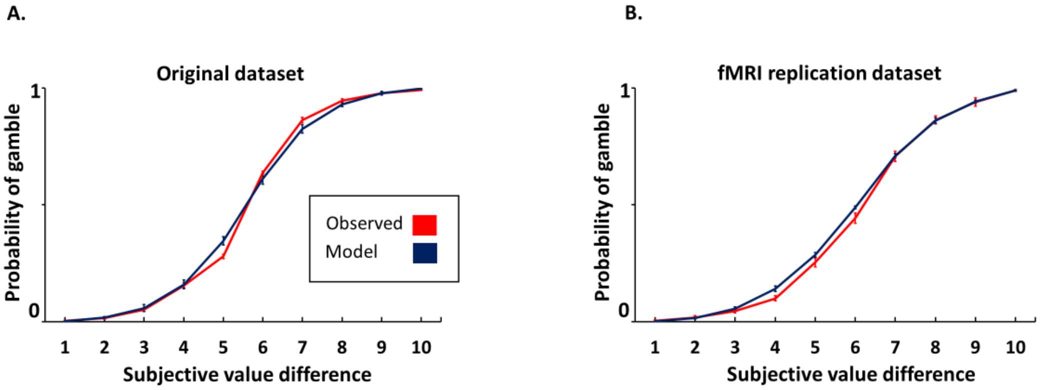


Figure S5. Prospect theory model simulations. Prospect theory model fits to the observed choice data for participants in the original (A) and fMRI replication (B) datasets. The probability of selecting to gamble during the baseline task is shown on the y-axis. The participants' subjective expected value differences between the risky gamble and the safe alternative are divided into 10 bins along the x-axis. The number 1 indicates those 10% of the observations where the difference is highest in favor of the safe alternative while 10 indicates those 10% of the observations where the difference is highest in favor of the risky choice. Error bars represent s.e.m computed across participants.

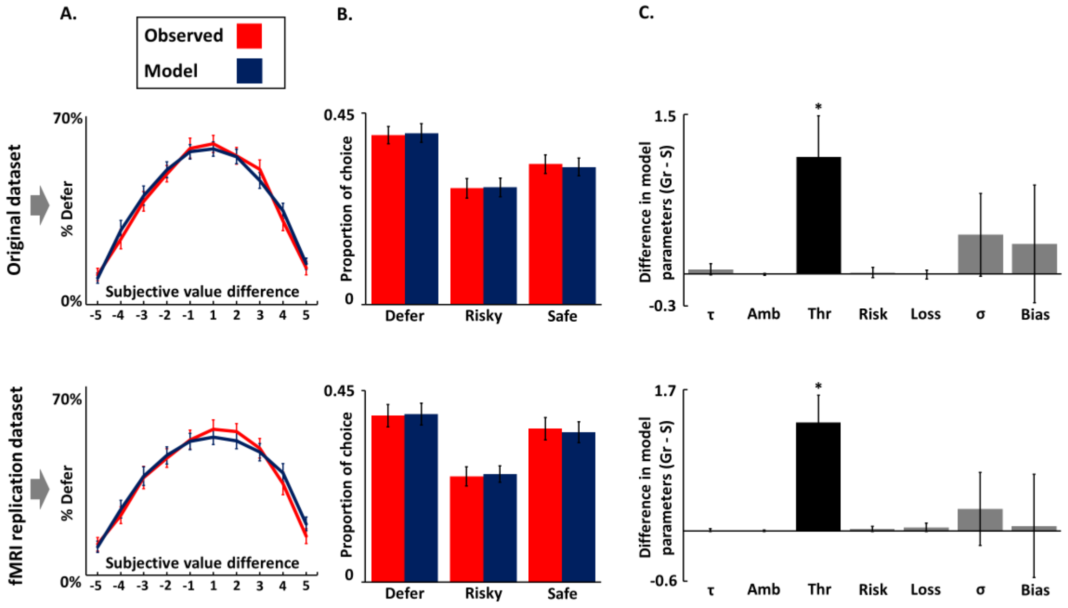


Figure S6. Computational modeling results depicted in Fig. 4 divided by dataset (full model). **A.** Model simulations (blue) vs. observed data (red) averaged across the Group and Self conditions. In both the original and fMRI replication datasets, the model is highly accurate in explaining the variance in the observed behavior. **B.** Model simulations of the average proportion of choices for each of the three alternative options (blue) compared to empirically observed choices (red). **C.** Differences in model parameter values in *Group* and *Self* trials. The *Group* trials, where subjects make decisions about taking responsibility for others, increases the deferral threshold such that a larger difference in subjective value is needed for participants to choose to lead. No other parameter is affected by the Group trials. Upper and lower panels represent data from the full model for the original and fMRI replication datasets respectively. For A. and B., error bars represent s.e.m., for C., errors bars represent s.d because they are obtained from a posterior distribution on the population level (see Supplementary Methods). * The posterior probability of a difference between the conditions =0.995 and 0.999 for the original and fMRI replication datasets, respectively. τ= stochasticity in the binary choice process, σ= noise in the readout of the utility, Amb=ambiguity preference measure, Thr= deferral threshold, Risk=risk preference measure, Loss = loss preference measure.

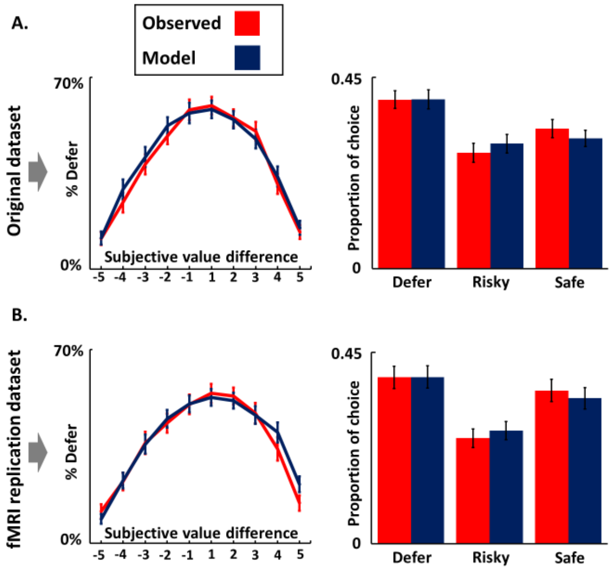


Figure S7. Supplementary computational results (restricted model). Model predictions and observed data for the original group (**A)** and fMRI replication group (**B**) from a version of our Delegation Model that restricted the PT parameters to be equal to those fit to choices made in the Baseline task. On the left, the percentage of choices to defer (averaged across Group and Self trials) as a function of the subjective value difference between the safe and risky options (10 bins). On the right, the average proportion of choices for each of the three alternative options across all trial types. For comparison with Table S2, the model fit measures (DIC) for this restricted model were 14407 and 16871 in the original and fMRI replication groups respectively.

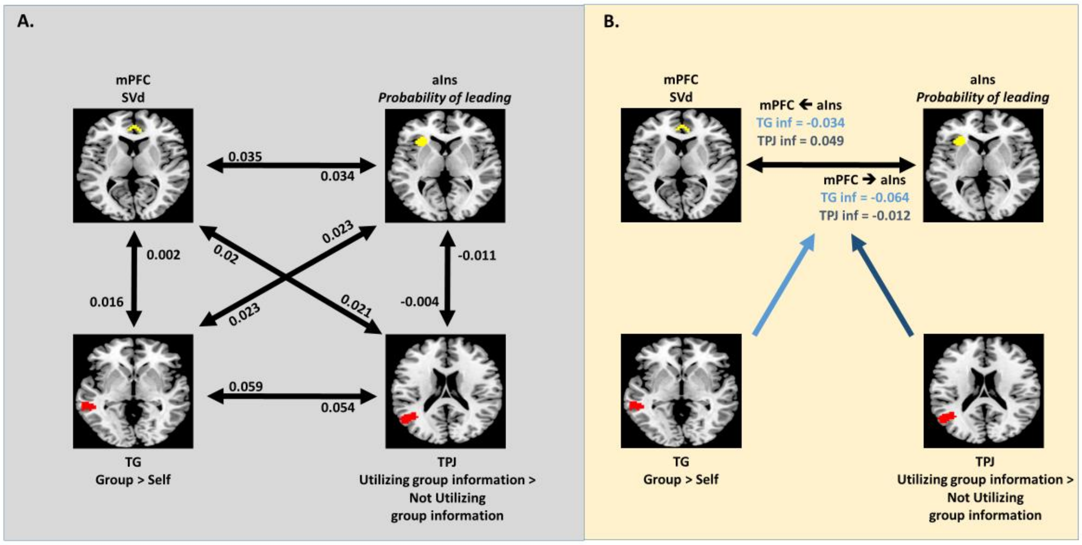


*Figure S8. DCM Group level estimates***.** Intrinsic connectivity parameters between nodes in the DCM network (from optimized model). **A.** Direct connections between regions in the network. **B.** Influence of activity in the TG (light blue) and TPJ (dark blue) on the connectivity between mPFC and aIns.

IV. Supplementary Tables

Table S1. Regression and correlation results.

S1A – Original group **Individual difference measure**	**Correlation with RA**	**Regression with RA as DV**	**Regression with LDBQ leadership score as DV**
Responsibility aversion	-	-	-0.94 (0.04)
Dictator game in-group	0.02 (0.89)	-0.00 (0.99)	-0.09 (0.71)
Dictator game out-group	0.11 (0.52)	0.06 (0.55)	-0.3 (0.32)
Affiliation in-group	-0.07 (0.68)	0.02 (0.83)	0.07 (0.78)
Affiliation out-group	-0.17 (0.30)	-0.05 (0.49)	-0.54 (0.12)
Performance Baseline task	-0.15 (0.38)	-0.02 (0.84)	-0.39 (0.36)
Performance Delegation task	-0.02 (0.90)	0.005 (0.95)	0.22 (0.38)
Individual's ambiguity preference	0.07 (0.66)	0.09 (0.22)	0.04 (0.89)
Individual's risk preference	-0.05 (0.75)	-0.02 (0.83)	-0.49 (0.31)
Individual's loss preference	0.00 (0.95)	-0.01 (0.92)	-0.91 (0.07)
Average peers risk preference	-0.05 (0.76)	0.04 (0.84)	0.50 (0.48)
Average peers loss preference	0.04 (0.80)	-0.12 (0.30)	-0.50 (0.17)

S1B – Replication group **Individual difference measure**	**Correlation with RA**	**Regression with RA as DV**	**Regression with composite leadership score as DV**
Responsibility aversion	-	-	-0.43 (0.02)
Dictator game in-group	0.27 (0.09)	0.45 (0.07)	-0.02 (0.94)
Dictator game out-group	0.34 (0.03)	-0.004 (0.99)	0.05 (0.83)
Affiliation in-group	0.04 (0.81)	-0.17 (0.42)	-0.001 (0.99)
Affiliation out-group	0.24 (0.13)	-0.03 (0.88)	-0.03 (0.89)
Performance Baseline task	0.02 (0.91)	0.30 (0.13)	-0.08 (0.70)
Performance Delegation task	-0.16 (0.31)	-0.13 (0.51)	0.37 (0.06)
Individual's ambiguity preference	0.28 (0.07)	0.16 (0.39)	0.19 (0.30)
Individual's risk preference	-0.02 (0.91)	0.07 (0.69)	-0.26 (0.14)
Individual's loss preference	0.18 (0.24)	0.26 (0.28)	0.01 (0.96)
Average peers risk preference	0.08 (0.63)	0.35 (0.20)	0.29 (0.27)
Average peers loss preference	0.10 (0.53)	-0.16 (0.57)	-0.15 (0.59)

Regression results with Responsibility Aversion (RA) or leadership scores as the dependent variable (DV) and direct correlations with RA for the original (A) and fMRI replication (B) groups. All correlation coefficients in column 1 represent Spearman's rho and the *p*-values in parentheses are derived from two-tailed tests. The regression coefficients in columns 2 and 3 are derived from ordinary least squares regressions after normalizing the regressors, and the *p*-values in parentheses are derived from two-tailed tests. For additional information on how each independent variable was obtained see Methods. In brief, risk and loss preferences were estimated via the PT model on the separate baseline task. Ambiguity preferences were obtained via a modified Ellsberg procedure. Social preferences (dictator and affiliation) were elicited from participants in a separate questionnaire. Performance represents the accumulated earnings over all events. For increased interpretability, all independent variables were normalized. Pro-sociality and democratic tendencies could, in principle, affect deferral behavior in the leadership condition because other participants' payoffs are affected by decisions in this condition. For this reason, pro-sociality might be related to a participant's responsibility aversion although it is theoretically not clear whether more prosocial participants will be more or less responsibility averse. We measured participants' pro-sociality in an independent dictator game task (see Methods 2.4). For the replication group we directly tested the more precise measure of responsibility aversion (bound shift). For robustness we additionally tested the regressions using a non-parametric approach (*75*). Using this approach, no significant association was found between any parameter in Table S1 and responsibility aversion (all p>0.19 uncorrected) and the strongest correlate of the leadership scores was responsibility aversion (beta=-0.89, *p*=0.06, beta=-0.45, p=0.02 and beta=-0.44, p=0.001, for the original group, replication group, and all subjects respectively). For significant correlations between the separate components of the composite leadership score and responsibility aversion see methods 2.3.3. See also **Fig. S2** for Bayesian evidence for null behavioral correlations with leadership scores.

Table S2. Model comparison for different versions of the Delegation Model.

Original Group:

	PT Variable	PT Constrained
OC Variable	**13073**	**13080**
OC Constrained	13103	13142

Replication Group:

	PT Variable	PT Constrained
OC Variable	**15343**	**15340**
OC Constrained	15371	15369

The tables in A) and B) show the relative model fits for four different parameterization combinations of the Delegation Model in the original and fMRI replication experiments, respectively. The label *Constrained* indicates that a parameter set was constrained to be equal across the Group and Self trials. The label *Variable* indicates that a parameter set was estimated separately in the two conditions. The quality of the model fits was assessed via a leave one out (LOO) approach based on Markov Chain Monte Carlo (MCMC) samples (*113*), where smaller values indicate a better fit. This approach represents a more accurate measure of model fit, after accounting for complexity than information criterion such as the Deviance Information Criterion (DIC) (*113*). The four model specifications differ in terms of whether or not the Prospect Theory (PT, see eq. 5) parameters used to estimate subjective values or the Optimal Categorization (OC, see eq. 7) parameters, including the deferral thresholds, were allowed to vary (i.e. estimated separately) across the Group and Self conditions. The bold text in the top row highlights the fact that model specifications allowing the OC parameters to vary across the Group and Self choices fit the data best. These model comparison results are consistent with the direct parameter comparisons reported in **Fig. 4C** and demonstrate that the difference between the conditions is driven by differences in OC deferral threshold parameter and not any of the PT parameters.

Table S3. Whole brain corrected contrasts used to identify ROI’s for DCM analysis.

Contrasts	**Brain Region**	**MNI**	t- value	*p*-value
Being in the context of responsibility Group>Self Across all events	Middle/Superior Temporal Gyrus (TG, BA 21/22)	-63 -39 -3	5.21	0.011
Group<Self	No significant activations			
Relying on the group’s decision* Letting the group decide > deciding alone. Across all events	Temporal Parietal Junction (TPJ, BA 39)	-45 -60 24	7.59	0.000005
Letting the group decide < deciding alone. Across all events	No significant activations			
Subjective value difference (*SVd*) Parametric regressor (Subjective value risky-safe option) across all events	Medial prefrontal cortex (mPFC, BA 32)	12 39 18	4.19	0.028
Probability of leading Parametric regressor across all events	Anterior Insula (aIns, BA 13)	-27 24 3	9.21	0.000

For clarity, this table reports only those regions selected for inclusion in our minimalistic brain network and DCM analysis. The complete set of whole-brain Family-Wise-Error (FWE) corrected results for each contrast at the cluster and voxel level are listed in **Tables S4** and **S5**. All regions survive cluster size correction for multiple comparisons across the whole brain, as determined by both parametric and non-parametric methods. Specifically, we initially used the default approach in SPM [FWE<0.05, initial cluster defining threshold $p < 0.001$ (*120*)] to define the regions of interest for our DCM. However, given recently published concerns related to cluster size correction (*121*), we also repeated the analysis and confirmed all of our univariate results using non-parametric permutation tests as implemented in the SnPM 13 toolbox [(http://warwick.ac.uk/snpm) with an initial cluster defining threshold of $p < 0.0001$ and 5000 permutations]. The peak-voxel t-values we report in all tables are from the original parametric analysis. DCM analyses require the definition of a model space (i.e. set of brain regions) and the choice of which regions is often based on both the current data and previous observations.(*80*) We follow this practice by selecting regions that are strongly linked to specific computations in both our current data and previous reports. Therefore, the most significant peak in our results was not always included in the model. Specifically, we opted not to use the occipital areas identified in the responsibility contrast because we believed that they may be more related to visual attention processes than responsibility per-se (*82*) and instead included the Middle/Superior Temporal Gyrus region that has been found to be related to the ability to distinguish between aspects of the self and of others(*83*, *84*). Similarly, given prior evidence for the neural representation of *SVd* signals(*29*, *30*, *85*), we selected the mPFC region for the *SVd* contrast. For simplicity, we selected the left hemisphere peak from the bilateral aIns activity. BA= Brodmann Area. * *related to* the informational advantage for deferring to others see supplementary section 10.2)

Table S4. Full list of activations for main contrasts.

Contrasts	Brain Region	MNI	Cluster size	t-value	*p*-value (FWE cluster size)
Being in the context of responsibility Group>Self	Occipital cortex; BA 19	12 -63 -6	174	7.52	0.00002
	Occipital cortex; BA 18	-12 -99 6	67	5.72	0.009
	Middle/Superior Temporal Gyrus (TG, BA 21/22)	-63 -39 -3	64	5.21	0.01
	Thalamus, Pulvinar	6 -30 -3	58	4.68	0.017
Group<Self	No significant activations				
Relying on the group's decision Letting the group decide > deciding alone. Across all events	Temporal Parietal Junction (TPJ, BA 39)	-45 -60 24	332	7.59	0.000005
	Dorsolateral prefrontal cortex (BA 8)	-18 30 45	216	6.62	0.0002
Letting the group decide < deciding alone. Across all events	No significant activations				
Subjective value difference (*SVd*) Parametric regressor (Subjective value risky-safe option) across all events	Cerebellum	0 -57 -36	1966	6.49	0
	Occipital cortex; BA 19	30 -81 12	1087	6.49	0
	Bi-lateral caudate	-9 12 -6	1055	9.53	0
	Inferior Frontal Gyrus (BA 46)	-39 42 12	218	6.85	0.000008
	Cingulate (BA 31)	-3 -30 36	156	5.57	0.0001
	Inferior Frontal Gyrus (BA 9)	-42 6 30	150	6.5	0.0002
	Inferior Frontal Gyrus (BA 46)	39 36 12	124	4.94	0.0006
	Cingulate (BA 24)	-3 6 24	86	5.84	0.005
	Inferior Parietal Lobe (BA 40)	48 -36 45	75	4.69	0.01

Subjective value difference (*SVd*) (continued)	Medial prefrontal cortex (mPFC, BA 32)	12 39 18	59	4.19	0.027
	Anterior insula (BA 13)	39 24 -3	59	4.95	0.027
	Inferior Frontal Gyrus (BA 9)	48 6 24	55	4.83	0.036
Probability of leading Parametric regressor across all events	Bilateral Anterior Insula (aIns, BA 13)	36 21 0	3637	9.42	0.000
	Fusiform Gyrus (BA 20)	36 -39 -21	3503	6.72	0.000
	Medial Frontal Gyrus (BA 9)	9 39 36	1477	7.77	0.000
	Inferior Parietal Lobe (BA 40)	-33 -48 36	699	5.98	0.000

This table reports all clusters surviving our primary analysis threshold of $p < 0.05$ FWE corrected across the whole brain at the cluster-level based on Random Field Theory within SPM12 (*122*). The initial cluster defining threshold was set to $p<0.001$(*120*). As noted in Table S3, we have subsequently confirmed all multiple comparison corrections using a non-parametric permutation approach. Due to the low spatial specificity of some extremely large clusters we additionally provide Table S5 using FWE correction on the voxel level.

Table S5. Activations surviving an FWE threshold of p<0.05 at the voxel level.

Contrasts FWE whole brain corrected on voxel level k>10	**Brain Region**	**MNI**	Cluster size	t-value	*p*-value FWE voxel level
Being in the context of responsibility Group>Self Across all events	Occipital cortex (BA 19)	12 -63 -6	27	7.52	0.00008
Relying on the group's decision Letting the group decide > deciding alone. Across all events	Temporal Parietal Junction (TPJ, BA 39)	-45 -60 24	35	7.59	0.0002
	Dorsolateral prefrontal cortex (BA 8)	-18 30 45	12	6.62	0.016
Subjective value difference (*SVd*) Parametric regressor (Subjective value risky-safe option) across all events	Left caudate	-9 12 -6	115	9.53	0.000
	Right caudate	9 12 -6 12 0 -6	121	8.56	0.000
	Thalamus	-3 -27 -6 6 -27 -6	33	8.50	0.000
	Dorsolateral Prefrontal cortex (Ba 10)	-39 42 12	20	6.85	0.0007
	Inferior Frontal Gyrus (BA 9)	-42 6 30	28	6.50	0.0021
	Occipital cortex (BA 19)	30 -81 12	25	6.49	0.0021
	Parietal Lobe (BA 7)	24 -66 42 15 -69 42	33	6.42	0.0027
	Occipital cortex (BA 18)	-27 -84 9 -30 -90 18	21	6.30	0.004
	Occipital cortex (BA 19)	-33 -75 24 -30 -84 27	19	6.11	0.0076
Probability of leading Parametric regressor across all events	Right Anterior Insula (aIns, BA 13)	36 21 0 27 21 -15 27 33 -12	189	9.42	0.000

Probability of leading (continued)	Left Anterior Insula (aIns, BA 13)	-27 24 3 -27 24 -12	116	9.21	0.000
	Left Posterior Insula (BA 13)	-33 15 27	19	8.06	0.00002
	Medial Frontal Gyrus (BA 9)	9 39 36 9 24 51	43	7.77	0.00004
	Middle Frontal Gyrus (BA 46)	42 27 21	35	7.59	0.00006
	Occipital cortex (BA 18)	-21 -87 -18 -12 -93 -6 -18 -99 -9	46	6.55	0.00176
	Right caudate	12 12 0 15 6 12	31	6.37	0.00324
	Posterior Cingulate (BA 23)	-3 -33 27	12	6.35	0.00341
	Occipital cortex (BA 18)	-9 -75 -6 -9 -72 6 -3 -75 0	14	6.13	0.007
	Posterior Cingulate (BA 31)	6 -39 36 9 -42 45	13	5.88	0.016

This table is supplementary to Table S4 and is provided because the presence of several extremely large clusters after cluster-level correction for multiple comparisons limits the spatial specificity of Table S4. The only difference between Tables S4 and the current Table S5 is that the correction for multiple comparisons was applied at the cluster versus voxel levels, respectively. All statistical inferences in the manuscript are derived from Table S4.

Table S6. DCM parameters significantly associated with individual variability in Responsibility aversion.

Regression Parameter	*Beta value*	*Significance**
TG influence on connection from mPFC to aIns	-1.47	$p < 0.0001$
TPJ differential influence on aIns (Group > Self)	1.35	$p < 0.0001$
Local activity in TPJ when deciding on your own	-0.75	$p < 0.0001$
Local activity in TG in the Group condition	-0.74	$p < 0.0001$
mPFC differential influence on aIns (Group > Self)	0.70	$p < 0.0001$
TG influence on connection from aIns to mPFC	1.06	$p < 0.0005$
Intrinsic connection from TG to mPFC	0.80	$p < 0.0005$
TPJ differential influence on TG (Group > Self)	-0.77	$p < 0.0005$
Local activity in TPJ when choosing to utilize group information	0.75	$p < 0.0005$
Intrinsic connection from mPFC to TG	-0.72	$p < 0.0005$
TPJ influence on connection from aIns to mPFC	0.72	$p < 0.0005$

As detailed in DCM methods section, we used our Self condition as a baseline in order to remove any confounds not specifically related to responsibility aversion and thus include in this analysis the deferential connectivity values (Group-Self). *The *p*-values listed represent uncorrected values, but all reported connectivity changes survive FWE Bonferroni-Holm correction for multiple comparisons at p <0.01.

Table S7: Parameter Recovery

		RECOVERED PARAMETERS	
Parameter	**Simulated Change**	**Mean**	**95% Credible interval**
Ambiguity (θ)	0	0	[-0.01; 0.01]
Risk (γ)	0	0	[-0.03; 0.03]
Loss (λ)	0	-0.01	[-0.06; 0.02]
Criterion bias *(b)*	0	0.3	[-0.20; 0.80]
Boundary criterion (κ)	1.2	1.13	[0.84; 1.41]
Softmax temp. (τ)	0	-0.01	[-0.02; 0.01]
SVd input noise (σ)	4	3.81	[3.39; 4.23]
Ambiguity (θ)	0	0	[-0.01; 0.01]
Risk (γ)	-0.1	-0.1	[-0.13; -0.07]
Loss (λ)	0	0	[-0.04; 0.04]
Criterion bias *(b)*	0	0	[-0.42; 0.39]
Boundary criterion (κ)	1.2	1.27	[1.00; 1.53]
Softmax temp. (τ)	0	0	[-0.01; 0.01]
SVd input noise (σ)	0	0.05	[-0.29; 0.39]
Ambiguity (θ)	0	0.01	[0.00; 0.02]
Risk (γ)	0	0.02	[-0.01; 0.05]
Loss (λ)	-0.63	-0.61	[-0.64; -0.57]
Criterion bias *(b)*	0	-0.19	[-0.56; 0.20]
Boundary criterion (κ)	1.2	1.27	[1.00; 1.54]
Softmax temp. (τ)	0	0	[-0.02; 0.01]
SVd input noise (σ)	0	0.11	[-0.23; 0.44]
Ambiguity (θ)	0.4	0.4	[0.39; 0.41]
Risk (γ)	0	0.03	[0.00; 0.06]
Loss (λ)	0	0	[-0.04; 0.04]
Criterion bias *(b)*	0	0.04	[-0.39; 0.49]
Boundary criterion (κ)	1.2	1.46	[1.18; 1.74]
Softmax temp. (τ)	0	0	[-0.01; 0.02]
SVd input noise (σ)	0	0.1	[-0.24; 0.42]
Ambiguity (θ)	0	0	[-0.01; 0.01]
Risk (γ)	0	0	[-0.03; 0.03]
Loss (λ)	0	-0.01	[-0.05; 0.02]
Criterion bias *(b)*	0	-0.09	[-0.49; 0.28]
Boundary criterion (κ)	1.2	1.19	[0.97; 1.42]
Softmax temp. (τ)	0.3	0.31	[0.28; 0.34]
SVd input noise (σ)	0	-0.1	[-0.40; 0.20]
Ambiguity (θ)	0	0	[-0.01; 0.01]
Risk (γ)	0	0.01	[-0.02; 0.04]
Loss (λ)	0	-0.01	[-0.04; 0.02]
Criterion bias *(b)*	1	1.29	[0.88; 1.68]
Boundary criterion (κ)	1.2	0.91	[0.67; 1.15]
Softmax temp. (τ)	0	0	[-0.01; 0.02]
SVd input noise (σ)	0	-0.23	[-0.52; 0.07]
Ambiguity (θ)	0	0.01	[0.00; 0.02]
Risk (γ)	0	0.01	[-0.02; 0.04]
Loss (λ)	0	0.02	[-0.01; 0.06]
Criterion bias *(b)*	0	-0.03	[-0.45; 0.38]
Boundary criterion (κ)	1.2	1.34	[1.08; 1.60]
Softmax temp. (τ)	0	0	[-0.01; 0.02]
SVd input noise (σ)	0	0.13	[-0.18; 0.44]
Ambiguity (θ)	0	0.01	[0.00; 0.02]
Risk (γ)	0	0.03	[0.00; 0.06]
Loss (λ)	0	0.04	[0.01; 0.08]
Criterion bias *(b)*	0	0.08	[-0.44; 0.59]
Boundary criterion (κ)	0	0.27	[-0.02; 0.53]
Softmax temp. (τ)	0	0	[-0.02; 0.01]
SVd input noise (σ)	4	4.4	[3.95; 4.84]

Table S7 continued		**Recovered Parameters**	
Parameter	**Simulated Change**	**Mean**	**95% Credible interval**
Ambiguity (θ)	0	0.01	[0.00; 0.02]
Risk (γ)	-0.1	-0.11	[-0.14; -0.08]
Loss (λ)	0	0.04	[0.00; 0.08]
Criterion bias *(b)*	0	0.15	[-0.26; 0.59]
Boundary criterion (κ)	0	-0.05	[-0.30; 0.20]
Softmax temp. (τ)	0	0.01	[0.00; 0.02]
SVd input noise (σ)	0	0.04	[-0.29; 0.37]
Ambiguity (θ)	0	0	[-0.01; 0.01]
Risk (γ)	0	0.03	[0.00; 0.06]
Loss (λ)	-0.63	-0.64	[-0.68; -0.61]
Criterion bias *(b)*	0	0.08	[-0.31; 0.47]
Boundary criterion (κ)	0	0.08	[-0.15; 0.33]
Softmax temp. (τ)	0	0.01	[0.00; 0.02]
SVd input noise (σ)	0	0.24	[-0.09; 0.58]
Ambiguity (θ)	0.4	0.4	[0.39; 0.41]
Risk (γ)	0	0.03	[0.00; 0.06]
Loss (λ)	0	0	[-0.03; 0.04]
Criterion bias *(b)*	0	-0.18	[-0.60; 0.25]
Boundary criterion (κ)	0	0.08	[-0.18; 0.35]
Softmax temp. (τ)	0	0	[-0.01; 0.02]
SVd input noise (σ)	0	0.01	[-0.30; 0.34]
Ambiguity (θ)	0	-0.01	[-0.02; 0.00]
Risk (γ)	0	0.03	[0.00; 0.06]
Loss (λ)	0	0	[-0.04; 0.04]
Criterion bias *(b)*	1	1.47	[1.06; 1.89]
Boundary criterion (κ)	0	0.22	[-0.02; 0.45]
Softmax temp. (τ)	0	0	[-0.01; 0.01]
SVd input noise (σ)	0	0.18	[-0.13; 0.49]

This table reports the results of a series of tests for how well our estimation procedures can recover known changes in the model's parameters. For each test, we constructed a set of choices from 20 simulated subjects performing our delegation task, changing a specific parameter or combination of the choice-generating parameters across Self and Group conditions. In total, we generated 20 choices for each trial and we had 280 trials (same number of trials as in our experiment). We then fit the model to the data from these simulated agents in order to determine if it could accurately and selectively identify the parameters that changed across conditions. In each simulation, the group-level parameters estimated from the real participants' data were used to generate choices in the Self condition. We first simulated six choice sets in which the deferral threshold plus one other model parameter changed in Group trials. In an additional set of six simulations, only a single model parameter changed when generating choices in Group versus Self trials. These results show that the model can recover known changes in parameter values. Critically, a selective change in the deferral threshold (i.e. our empirical result) was never observed when the true change in the choice generating process was driven by another model parameter.

V. References and Notes:

1. B. Bass, R. Bass, *The Bass Handbook of Leadership: Theory, Research, and Managerial Applications* (Free Press, New York, 2009).

2. S. M. Fleming, N. D. Daw, Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychol. Rev.* **124**, 91–114 (2017). doi:10.1037/rev0000045 Medline

3. C. Summerfield, K. Tsetsos, Do humans make good decisions? *Trends Cogn. Sci.* **19**, 27–34 (2015). doi:10.1016/j.tics.2014.11.005 Medline

4. B. De Martino, S. Bobadilla-Suarez, T. Nouguchi, T. Sharot, B. C. Love, Social information is integrated into value and confidence judgments according to its reliability. *J. Neurosci.* **37**, 6066–6074 (2017). doi:10.1523/JNEUROSCI.3880-16.2017 Medline

5. B. Bartling, E. Fehr, H. Herz, The intrinsic value of decision rights. *Econometrica* **82**, 2005–2039 (2014). doi:10.3982/ECTA11573

6. E. Fehr, H. Herz, T. Wilkening, The lure of authority: Motivation and incentive effects of power. *Am. Econ. Rev.* **103**, 1325–1359 (2013). doi:10.1257/aer.103.4.1325

7. C. Eckel, P. Grossman, Managing diversity by creating team identity. *J. Econ. Behav. Organ.* **58**, 371–392 (2005). doi:10.1016/j.jebo.2004.01.003

8. J. Faria, J. Dyer, C. Tosh, J. Krause, Leadership and social information use in human crowds. *Anim. Behav.* **79**, 895–901 (2010). doi:10.1016/j.anbehav.2009.12.039

9. A. W. Halpin, B. J. Winer, in *Leader Behavior: Its Description and Measurement*, R. M. Stogdill, A. E. Coons, Eds. (Ohio State Univ., 1957).

10. R. Blake, J. Mouton, A. Bidwell, Managerial grid. *Adv. Manag.–Office Exec.* **1**, 12–15 (1962).

11. E. A. Fleishman, M. D. Mumford, S. J. Zaccaro, K. Y. Levin, A. L. Korotkin, M. B. Hein, Taxonomic efforts in the description of leader behavior: A synthesis and functional interpretation. *Leadersh. Q.* **2**, 245–287 (1992). doi:10.1016/1048-9843(91)90016-U

12. C. S. Burke, K. C. Stagl, C. Klein, G. F. Goodwin, E. Salas, S. M. Halpin, What type of leadership behaviors are functional in teams? A meta-analysis. *Leadersh. Q.* **17**, 288–307 (2006). doi:10.1016/j.leaqua.2006.02.007

13. T. A. Judge, R. F. Piccolo, R. Ilies, The forgotten ones? The validity of consideration and initiating structure in leadership research. *J. Appl. Psychol.* **89**, 36–51 (2004). doi:10.1037/0021-9010.89.1.36 Medline

14. G. Charness, M. Jackson, The role of responsibility in strategic risk-taking. *J. Econ. Behav. Organ.* **69**, 241–247 (2009). doi:10.1016/j.jebo.2008.10.006

15. J. Pahlke, S. Strasser, F. Vieider, Responsibility effects in decision making under risk. *J. Risk Uncertain.* **51**, 125–146 (2015). doi:10.1007/s11166-015-9223-6

16. D. M. Green, J. A. Swets, *Signal Detection Theory and Psychophysics* (Wiley, 1966).

17. A. T. Qamar, R. J. Cotton, R. G. George, J. M. Beck, E. Prezhdo, A. Laudano, A. S. Tolias, W. J. Ma, Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 20332–20337 (2013). doi:10.1073/pnas.1219756110 Medline

18. A. Tversky, D. Kahneman, Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertain.* **5**, 297–323 (1992). doi:10.1007/BF00122574

19. C. D. Frith, U. Frith, Mechanisms of social cognition. *Annu. Rev. Psychol.* **63**, 287–313 (2012). doi:10.1146/annurev-psych-120710-100449 Medline

20. Y.-L. Boureau, P. Sokol-Hessner, N. D. Daw, Deciding how to decide: Self-control and meta-decision making. *Trends Cogn. Sci.* **19**, 700–710 (2015). doi:10.1016/j.tics.2015.08.013 Medline

21. B. J. Avolio, B. M. Bass, D. I. Jung, Re-examining the components of transformational and transactional leadership using the Multifactor Leadership. *J. Occup. Organ. Psychol.* **72**, 441–462 (1999). doi:10.1348/096317999166789

22. P. Balthazard, D. Waldman, R. W. Thatcher, S. T. Hannah, Differentiating transformational and non-transformational leaders on the basis of neurological imaging. *Leadersh. Q.* **23**, 244–258 (2012). doi:10.1016/j.leaqua.2011.08.002

23. S. Einarsen, M. S. Aasland, A. Skogstad, Destructive leadership behaviour: A definition and conceptual model. *Leadersh. Q.* **18**, 207–216 (2007). doi:10.1016/j.leaqua.2007.03.002

24. J. Antonakis, D. V. Day, B. Schyns, Leadership and individual differences: At the cusp of a renaissance. *Leadersh. Q.* **23**, 643–650 (2012). doi:10.1016/j.leaqua.2012.05.002

25. R. E. Boyatzis, K. Rochford, A. I. Jack, Antagonistic neural networks underlying differentiated leadership roles. *Front. Hum. Neurosci.* **8**, 114 (2014). doi:10.3389/fnhum.2014.00114 Medline

26. J. Jiang, C. Chen, B. Dai, G. Shi, G. Ding, L. Liu, C. Lu, Leader emergence through interpersonal neural synchronization. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 4274–4279 (2015). doi:10.1073/pnas.1422930112 Medline

27. S. L. Bressler, V. Menon, Large-scale brain networks in cognition: Emerging methods and principles. *Trends Cogn. Sci.* **14**, 277–290 (2010). doi:10.1016/j.tics.2010.04.004 Medline

28. B. Li, J. Daunizeau, K. E. Stephan, W. Penny, D. Hu, K. Friston, Generalised filtering and stochastic DCM for fMRI. *Neuroimage* **58**, 442–457 (2011). doi:10.1016/j.neuroimage.2011.01.085 Medline

29. O. Bartra, J. T. McGuire, J. W. Kable, The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage* **76**, 412–427 (2013). doi:10.1016/j.neuroimage.2013.02.063 Medline

30. J. A. Clithero, A. Rangel, Informatic parcellation of the network involved in the computation of subjective value. *Soc. Cogn. Affect. Neurosci.* **9**, 1289–1302 (2014). doi:10.1093/scan/nst106 Medline

31. R. M. Carter, S. A. Huettel, A nexus model of the temporal-parietal junction. *Trends Cogn. Sci.* **17**, 328–336 (2013). doi:10.1016/j.tics.2013.05.007 Medline

32. V. Vroom, P. Yetton, *Leadership and Decision-Making* (Univ. Pittsburgh Press, 1973).

33. L. A. Leotti, M. R. Delgado, The value of exercising control over monetary gains and losses. *Psychol. Sci.* **25**, 596–604 (2014). doi:10.1177/0956797613514589 Medline

34. Consider the analogy with the intuitive notion of risk aversion, which lacked a precise theoretical underpinning before Arrow and Pratt precisely defined it in terms of the concavity of an individual's utility function, thereby opening the door for theoretical modeling and the precise interpretation of empirical measures of risk aversion. Without these foundations, progress in understanding the concept of risk aversion would have been seriously impeded.

35. J. Dickinson, Employees' preferences for the bases of pay differentials. *Employee Relat.* **28**, 164–183 (2006). doi:10.1108/01425450610639383

36. I. Krajbich, B. Bartling, T. Hare, E. Fehr, Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nat. Commun.* **6**, 7455 (2015). doi:10.1038/ncomms8455 Medline

37. G. Bolton, A. Ockenfels, J. Stauf, Social responsibility promotes conservative risk behavior. *Eur. Econ. Rev.* **74**, 109–127 (2015). doi:10.1016/j.euroecorev.2014.10.002

38. F. Song, Trust and reciprocity behavior and behavioral forecasts: Individuals versus group-representatives. *Games Econ. Behav.* **62**, 675–696 (2008). doi:10.1016/j.geb.2007.06.002

39. P. Molenberghs, G. Prochilo, N. K. Steffens, H. Zacher, S. A. Haslam, The neuroscience of inspirational leadership: The importance of collective-oriented language and shared group membership. *J. Manage.* **43**, 1–27 (2015).

40. S. Chow, H. Wang, J. Shao, *Sample Size Calculations in Clinical Research* (Chapman and Hall, ed. 2, 2007).

41. G. Yukl, A. Gordon, T. Taber, A hierarchical taxonomy of leadership behavior: Integrating a half century of behavior research. *J. Leadersh. Organ. Stud.* **9**, 15–32 (2002). doi:10.1177/107179190200900102

42. R. Hogan, R. Kaiser, What we know about leadership. *Rev. Gen. Psychol.* **9**, 169–180 (2005). doi:10.1037/1089-2680.9.2.169

43. D. Ellsberg, Risk, ambiguity, and the Savage axioms. *Q. J. Econ.* **75**, 643–669 (1961). doi:10.2307/1884324

44. I. Levy, J. Snell, A. J. Nelson, A. Rustichini, P. W. Glimcher, Neural representation of subjective value under risk and ambiguity. *J. Neurophysiol.* **103**, 1036–1047 (2010). doi:10.1152/jn.00853.2009 Medline

45. P. C. Trimmer, A. I. Houston, J. A. R. Marshall, M. T. Mendl, E. S. Paul, J. M. McNamara, Decision-making under uncertainty: Biases and Bayesians. *Anim. Cogn.* **14**, 465–476 (2011). doi:10.1007/s10071-011-0387-4 Medline

46. A. E. Colbert, T. A. Judge, D. Choi, G. Wang, Assessing the trait theory of leadership using self and observer ratings of personality: The mediating role of contributions to group success. *Leadersh. Q.* **23**, 670–685 (2012). doi:10.1016/j.leaqua.2012.03.004

47. M. Bazerman, *The Power of Noticing: What the Best Leaders See* (Simon and Schuster, 2014).

48. K. L. Cullen-Lester, F. J. Yammarino, Collective and network approaches to leadership: Special issue introduction. *Leadersh. Q.* **27**, 173–180 (2016). doi:10.1016/j.leaqua.2016.02.001

49. J. Antonakis, B. J. Avolio, N. Sivasubramaniam, Context and leadership: An examination of the nine-factor full-range leadership theory using the Multifactor Leadership Questionnaire. *Leadersh. Q.* **14**, 261–295 (2003). doi:10.1016/S1048-9843(03)00030-4

50. T. A. Judge, R. F. Piccolo, Transformational and transactional leadership: A meta-analytic test of their relative validity. *J. Appl. Psychol.* **89**, 755–768 (2004). doi:10.1037/0021-9010.89.5.755 Medline

51. M. B. Eberly, M. D. Johnson, M. Hernandez, B. J. Avolio, An integrative process model of leadership: Examining loci, mechanisms, and event cycles. *Am. Psychol.* **68**, 427–443 (2013). doi:10.1037/a0032244 Medline

52. G. Graen, M. Uhl-Bien, Relationship-based approach to leadership: Development of leader-member exchange (LMX) theory of leadership over 25 years: Applying a multi-level multi-domain. *Leadersh. Q.* **6**, 219–247 (1995). doi:10.1016/1048-9843(95)90036-5

53. F. Fiedler, A contingency model of leadership effectiveness. *Adv. Exp. Soc. Psychol.* **1**, 149–190 (1964). doi:10.1016/S0065-2601(08)60051-9

54. J. Conger, R. Kanungo, *Charismatic Leadership in Organizations* (Sage Publications, Thousand Oaks, CA, 1998).

55. R. G. Lord, D. J. Brown, *Leadership Processes and Follower Self-Identity* (Erlbaum, Mahwah, NJ, 2004).

56. A. H. Eagly, M. C. Johannesen-Schmidt, M. L. van Engen, Transformational, transactional, and laissez-faire leadership styles: A meta-analysis comparing women and men. *Psychol. Bull.* **129**, 569–591 (2003). doi:10.1037/0033-2909.129.4.569 Medline

57. J. E. Bono, T. A. Judge, Personality and transformational and transactional leadership: A meta-analysis. *J. Appl. Psychol.* **89**, 901–910 (2004). doi:10.1037/0021-9010.89.5.901 Medline

58. N. Lee, C. Senior, M. Butler, Leadership research and cognitive neuroscience: The state of this union. *Leadersh. Q.* **23**, 213–218 (2012). doi:10.1016/j.leaqua.2011.08.001

59. S. T. Hannah, P. A. Balthazard, D. A. Waldman, P. L. Jennings, R. W. Thatcher, The psychological and neurological bases of leader self-complexity and effects on adaptive decision-making. *J. Appl. Psychol.* **98**, 393–411 (2013). doi:10.1037/a0032257 Medline

60. R. E. Boyatzis, A. M. Passarelli, K. Koenig, M. Lowe, B. Mathew, J. K. Stoller, M. Phillips, Examination of the neural substrates activated in memories of experiences with resonant and dissonant leaders. *Leadersh. Q.* **23**, 259–272 (2012). doi:10.1016/j.leaqua.2011.08.003

61. M. T. Fairhurst, P. Janata, P. E. Keller, Leading the follower: An fMRI investigation of dynamic cooperativity and leader-follower strategies in synchronization with an adaptive virtual partner. *Neuroimage* **84**, 688–697 (2014). doi:10.1016/j.neuroimage.2013.09.027 Medline

62. V. H. Vroom, A. G. Jago, The role of the situation in leadership. *Am. Psychol.* **62**, 17–24 (2007). doi:10.1037/0003-066X.62.1.17 Medline

63. C. Tabernero, M. J. Chambel, L. Curral, J. M. Arana, The role of task-oriented versus relationship-oriented

leadership on normative contract and group performance. *Soc. Behav. Personal.* **37**, 1391–1404 (2009). doi:10.2224/sbp.2009.37.10.1391

64. R. Blake, A. McCanse, *Leadership Dilemmas–Grid Solutions* (Gulf Professional Publishing, 1991).

65. J. Guyot, Management training and post-industrial apologetics. *Calif. Manage. Rev.* **20**, 84–93 (1978). doi:10.2307/41164786

66. M. Chemers, Leadership research and theory: A functional integration. *Group Dyn.* **4**, 27–43 (2000). doi:10.1037/1089-2699.4.1.27

67. C. Engel, Dictator games: A meta study. *Exp. Econ.* **14**, 583–610 (2011). doi:10.1007/s10683-011-9283-7

68. D. Kahneman, A. Tversky, Prospect theory: An analysis of decision under risk. *Econometrica* **47**, 263–292 (1979). doi:10.2307/1914185

69. H. Nilsson, J. Rieskamp, E.-J. Wagenmakers, Hierarchical Bayesian parameter estimation for cumulative prospect theory. *J. Math. Psychol.* **55**, 84–93 (2011). doi:10.1016/j.jmp.2010.08.006

70. J. Kruschke, *Doing Bayesian Data Analysis: A Tutorial With R, JAGS, and Stan* (Academic Press, ed. 2, 2014).

71. M. Plummer, in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, Technische Universität Wien, Vienna, Austria, 20 to 22 March 2003.

72. L. Savage, *The Foundations of Statistics* (Wiley, 1972).

73. A. Wheeler, A. Ganji, V. Krishnan, B. Thurow, *Introduction to Engineering Experimentation* (Prentice Hall, 1996).

74. R. Thompson, A note on restricted maximum likelihood estimation with an alternative outlier model. *J. R. Stat. Soc. B* **47**, 53–55 (1985).

5. D. Birkes, Y. Dodge, *Alternative Methods of Regression* (Wiley, 2011).

76. G. H. Glover, T. Q. Li, D. Ress, Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. *Magn. Reson. Med.* **44**, 162–167 (2000). doi:10.1002/1522-2594(200007)44:1<162:AID-MRM23>3.0.CO;2-E Medline

77. A. K. Harvey, K. T. S. Pattinson, J. C. W. Brooks, S. D. Mayhew, M. Jenkinson, R. G. Wise, Brainstem functional magnetic resonance imaging: Disentangling signal from physiological noise. *J. Magn. Reson. Imaging* **28**, 1337–1344 (2008). doi:10.1002/jmri.21623 Medline

78. C. Hutton, O. Josephs, J. Stadler, E. Featherstone, A. Reid, O. Speck, J. Bernarding, N. Weiskopf, The impact of physiological noise correction on fMRI at 7 T. *Neuroimage* **57**, 101–112 (2011). doi:10.1016/j.neuroimage.2011.04.018 Medline

79. K. E. Stephan, L. Kasper, L. M. Harrison, J. Daunizeau, H. E. M. den Ouden, M. Breakspear, K. J. Friston, Nonlinear dynamic causal models for fMRI. *Neuroimage* **42**, 649–662 (2008). doi:10.1016/j.neuroimage.2008.04.262 Medline

80. K. E. Stephan, W. D. Penny, R. J. Moran, H. E. M. den Ouden, J. Daunizeau, K. J. Friston, Ten simple rules for dynamic causal modeling. *Neuroimage* **49**, 3099–3109 (2010). doi:10.1016/j.neuroimage.2009.11.015 Medline

81. T. A. Hare, W. Schultz, C. F. Camerer, J. P. O'Doherty, A. Rangel, Transformation of stimulus value signals into motor commands during simple choice. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 18120–18125 (2011). doi:10.1073/pnas.1109322108 Medline

82. S. Kastner, L. G. Ungerleider, Mechanisms of visual attention in the human cortex. *Annu. Rev. Neurosci.* **23**, 315–341 (2000). doi:10.1146/annurev.neuro.23.1.315 Medline

83. C. D. Frith, U. Frith, Interacting minds—A biological basis. *Science* **286**, 1692–1695 (1999). doi:10.1126/science.286.5445.1692 Medline

84. V. Goel, J. Grafman, N. Sadato, M. Hallett, Modeling other minds. *Neuroreport* **6**, 1741–1746 (1995). doi:10.1097/00001756-199509000-00009 Medline

85. T. A. Hare, J. O'Doherty, C. F. Camerer, W. Schultz, A. Rangel, Dissociating the role of the orbitofrontal cortex

and the striatum in the computation of goal values and prediction errors. *J. Neurosci.* **28**, 5623–5630 (2008). doi:10.1523/JNEUROSCI.1309-08.2008 Medline

86. S. Rudorf, T. A. Hare, Interactions between dorsolateral and ventromedial prefrontal cortex underlie context-dependent stimulus valuation in goal-directed choice. *J. Neurosci.* **34**, 15988–15996 (2014). doi:10.1523/JNEUROSCI.3192-14.2014 Medline

87. K. Friston, W. Penny, Post hoc Bayesian model selection. *Neuroimage* **56**, 2089–2099 (2011). doi:10.1016/j.neuroimage.2011.03.062 Medline

88. M. J. Rosa, K. Friston, W. Penny, Post-hoc selection of dynamic causal models. *J. Neurosci. Methods* **208**, 66–78 (2012). doi:10.1016/j.jneumeth.2012.04.013 Medline

89. J. Friedman, T. Hastie, R. Tibshirani, *The Elements of Statistical Learning* (Springer, ed. 1, 2001).

90. A. Smith, B. Bernheim, C. Camerer, A. Rangel, "Neural activity reveals preferences without choices" (NBER Working Papers 19270, National Bureau of Economic Research, Cambridge, MA, 2013).

91. R. Tibshirani, Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).

92. H. Zou, T. Hastie, Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* **67**, 301–320 (2005). doi:10.1111/j.1467-9868.2005.00503.x

93. T. A. Hare, S. Hakimi, A. Rangel, Activity in dlPFC and its effective connectivity to vmPFC are associated with temporal discounting. *Front. Neurosci.* **8**, 50 (2014). doi:10.3389/fnins.2014.00050 Medline

94. R. Kass, A. Raftery, Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995). doi:10.1080/01621459.1995.10476572

95. K. Burnham, D. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (Springer, 2003).

96. J. Surowiecki, *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations* (Doubleday, 2004).

97. B. Bahrami, K. Olsen, P. E. Latham, A. Roepstorff, G. Rees, C. D. Frith, Optimally interacting minds. *Science* **329**, 1081–1085 (2010). doi:10.1126/science.1185718 Medline

98. P. Aghion, J. Tirole, Formal and real authority in organizations. *J. Polit. Econ.* **105**, 1–29 (1997). doi:10.1086/262063

99. O. John, S. Srivastava, in *Handbook of Personality: Theory and Research*, L. A. Pervin, O. P. John Eds. (Guilford Press, 1999), vol. 2, pp. 102–138.

100. G. Loomes, R. Sugden, Regret theory: An alternative theory of rational choice under uncertainty. *Econ. J. (Lond.)* **92**, 805–824 (1982). doi:10.2307/2232669

101. G. Charness, M. Dufwenberg, Promises and Partnership, Promises and partnership. *Econometrica* **74**, 1579–1601 (2006). doi:10.1111/j.1468-0262.2006.00719.x

102. G. Coricelli, H. D. Critchley, M. Joffily, J. P. O'Doherty, A. Sirigu, R. J. Dolan, Regret and its avoidance: A neuroimaging study of choice behavior. *Nat. Neurosci.* **8**, 1255–1262 (2005). doi:10.1038/nn1514 Medline

103. J. M. Leonhardt, L. R. Keller, C. Pechmann, Avoiding the risk of responsibility by seeking uncertainty: Responsibility aversion and preference for indirect agency when choosing for others. *J. Consum. Psychol.* **21**, 405–413 (2011). doi:10.1016/j.jcps.2011.01.001

104. M. Zeelenberg, J. Beattie, Consequences of regret aversion 2: Additional evidence for effects of feedback on decision making. *Organ. Behav. Hum. Decis. Process.* **72**, 63–78 (1997). doi:10.1006/obhd.1997.2730

105. 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). doi:10.1016/j.neuron.2011.02.056 Medline

106. R. F. Baumeister, A. M. Stillwell, T. F. Heatherton, Guilt: An interpersonal approach. *Psychol. Bull.* **115**, 243–267 (1994). doi:10.1037/0033-2909.115.2.243 Medline

107. T. Dohmen, A. Falk, D. Huffman, U. Sunde, J. Schupp, G. G. Wagner, Individual risk attitudes: Measurement, determinants, and behavioral consequences. *J. Eur. Econ. Assoc.* **9**, 522–550 (2011). doi:10.1111/j.1542-

4774.2011.01015.x

108. A. Reynaud, S. Couture, Stability of risk preference measures: Results from a field experiment on French farmers. *Theory Decis.* **73**, 203–221 (2012). doi:10.1007/s11238-012-9296-5

109. O. Andersson, H. Holm, J. Tyran, E. Wengström, Deciding for others reduces loss aversion. *Manage. Sci.* **62**, 29–36 (2014).

110. J. Pahlke, S. Strasser, F. M. Vieider, Risk-taking for others under accountability. *Econ. Lett.* **114**, 102–105 (2012). doi:10.1016/j.econlet.2011.09.037

111. F. M. Vieider, C. Villegas-Palacio, P. Martinsson, M. Mejía, Risk taking for oneself and others: A structural model approach. *Econ. Inq.* **54**, 879–894 (2015). doi:10.1111/ecin.12290

112. N. Barberis, Thirty years of prospect theory in economics: A review and assessment. *J. Econ. Perspect.* **27**, 173–196 (2013). doi:10.1257/jep.27.1.173

113. A. Vehtari, A. Gelman, J. Gabry, Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27**, 1413–1432 (2017). doi:10.1007/s11222-016-9696-4

114. M. Schurz, J. Radua, M. Aichhorn, F. Richlan, J. Perner, Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neurosci. Biobehav. Rev.* **42**, 9–34 (2014). doi:10.1016/j.neubiorev.2014.01.009 Medline

115. R. E. Huber, V. Klucharev, J. Rieskamp, Neural correlates of informational cascades: Brain mechanisms of social influence on belief updating. *Soc. Cogn. Affect. Neurosci.* **10**, 589–597 (2015). doi:10.1093/scan/nsu090 Medline

116. S. Guionnet, J. Nadel, E. Bertasi, M. Sperduti, P. Delaveau, P. Fossati, Reciprocal imitation: Toward a neural basis of social interaction. *Cereb. Cortex* **22**, 971–978 (2012). doi:10.1093/cercor/bhr177 Medline

117. C. Feng, G. Deshpande, C. Liu, R. Gu, Y.-J. Luo, F. Krueger, Diffusion of responsibility attenuates altruistic punishment: A functional magnetic resonance imaging effective connectivity study. *Hum. Brain Mapp.* **37**, 663–677 (2016). doi:10.1002/hbm.23057 Medline

118. U. Toelch, D. R. Bach, R. J. Dolan, The neural underpinnings of an optimal exploitation of social information under uncertainty. *Soc. Cogn. Affect. Neurosci.* **9**, 1746–1753 (2014). doi:10.1093/scan/nst173 Medline

119. F. Kurth, K. Zilles, P. T. Fox, A. R. Laird, S. B. Eickhoff, A link between the systems: Functional differentiation and integration within the human insula revealed by meta-analysis. *Brain Struct. Funct.* **214**, 519–534 (2010). doi:10.1007/s00429-010-0255-z Medline

120. C.-W. Woo, A. Krishnan, T. D. Wager, Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. *Neuroimage* **91**, 412–419 (2014). doi:10.1016/j.neuroimage.2013.12.058 Medline

121. A. Eklund, T. E. Nichols, H. Knutsson, Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7900–7905 (2016). doi:10.1073/pnas.1602413113 Medline

122. K. J. Friston, K. J. Worsley, R. S. Frackowiak, J. C. Mazziotta, A. C. Evans, Assessing the significance of focal activations using their spatial extent. *Hum. Brain Mapp.* **1**, 210–220 (1994). doi:10.1002/hbm.460010306 Medline

VI. Appendix S1. Example instructions for the Baseline task.

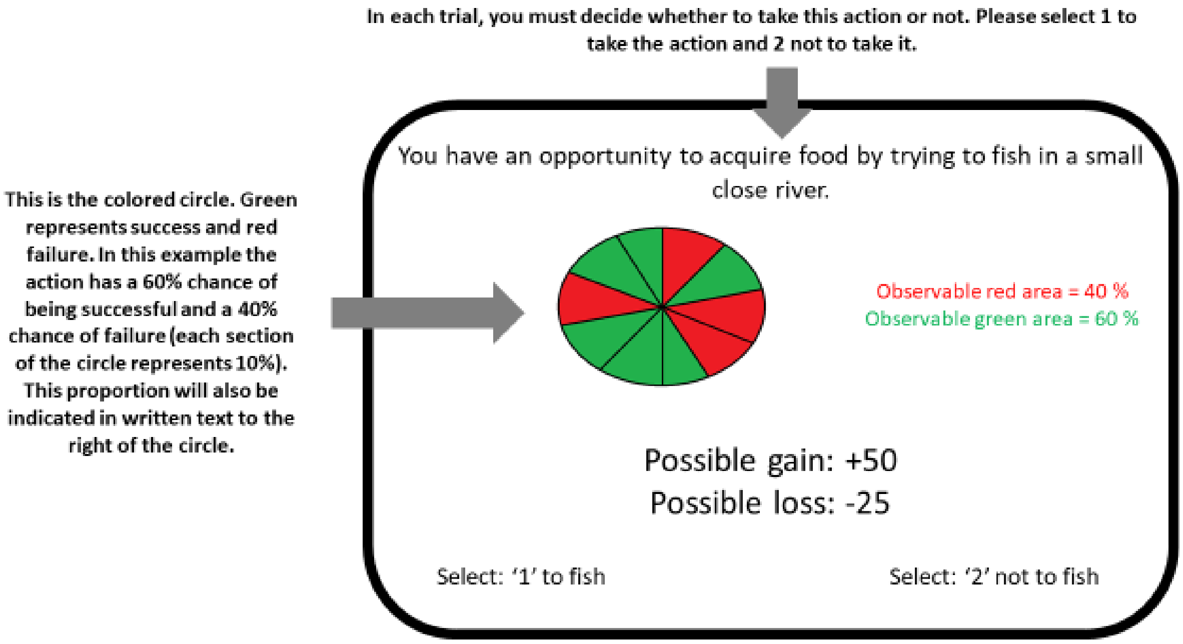
Welcome to the Econ-Lab!

You are now participating in an experiment at the Department of Economics, University of Zurich. Please read the following instructions carefully for at least 10 minutes. Fully understanding the instructions will allow you to perform better on the task. We will go through a training session before the task, but please take the time to understand the instructions fully. If you have any questions, please raise your hand and an experimenter will come over to help.

<u>Task:</u>

Imagine yourself racing through a jungle. You will be faced with a series of actions that can affect your survival (e.g. crossing a river, finding food etc.). You can always decide whether to perform or not perform the action. Each action has a positive consequence if it succeeds but a negative consequence if it fails. These values will be clearly marked on the screen (see example screen below). The probability of each action to succeed or fail (that is, how likely it is to work) will also be presented on the screen in the form of colored portions of a circle (see example below). The color green represents success and red represents failure. Thus the more green the circle is, the better the chances this action will succeed. Each section of the circle represents 10% of the total (see picture below).

Example trial screen:



In each trial, you must decide whether to take this action or not. Please select 1 to take the action and 2 not to take it.

This is the colored circle. Green represents success and red failure. In this example the action has a 60% chance of being successful and a 40% chance of failure (each section of the circle represents 10%). This proportion will also be indicated in written text to the right of the circle.

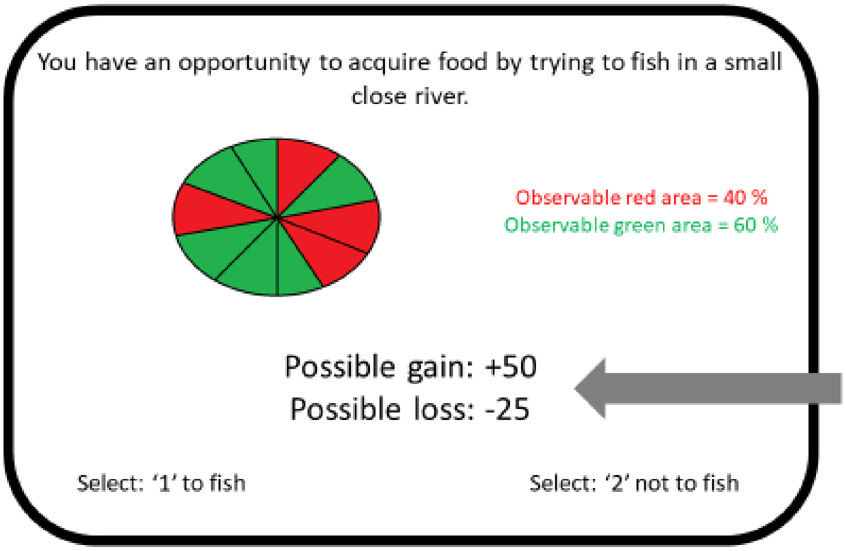
You have an opportunity to acquire food by trying to fish in a small close river.

Observable red area = 40 %
Observable green area = 60 %

Possible gain: +50
Possible loss: -25

Select: '1' to fish

Select: '2' not to fish



You have an opportunity to acquire food by trying to fish in a small close river.

Observable red area = 40 %
Observable green area = 60 %

Possible gain: +50
Possible loss: -25

Select: ‘1’ to fish

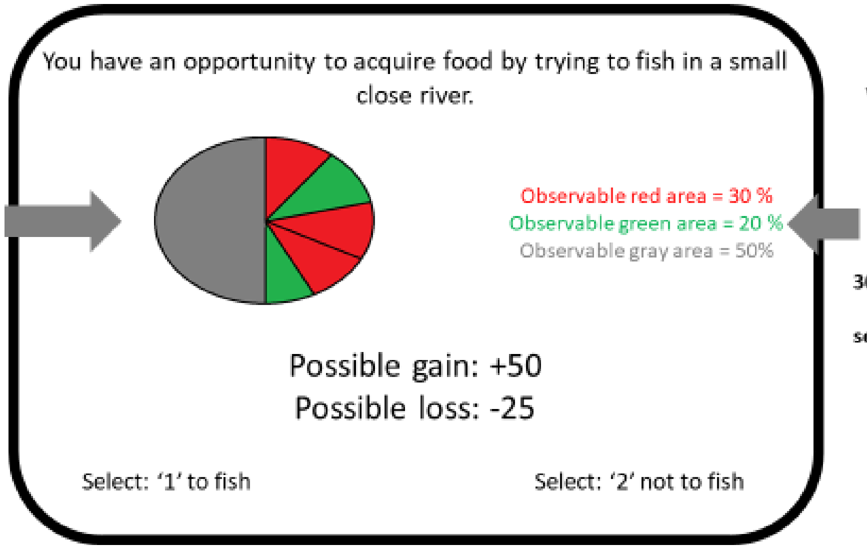
Select: ‘2’ not to fish

In this example trial: if you decide to take the action and it is successful, you will earn 50 points. However, if you decide to take the action and it did not succeed you will lose 25 points. You can also choose not to take the action at all, in which case you will neither earn nor lose for this trial (outcome = 0).

If you decide to take the action, the computer will determine if the action was successful according to the proportion of green and red color in the circle. This is always a gamble however, and success or failure are never insured. For example if the probability of success is 60% (as in the example above) this means that this action would succeed 6 out of 10 times but will fail 4 out of 10. As an illustration, you can imagine the computer is throwing a 10 point dice after each trial you selected to act; in the example above, if the dice throw resulted in a number between 1-6 you would win 50 but if it happened to be 7-10 you will lose 25. If you decide not to take any action you will not win or lose anything for this trial (i.e. your outcome for this trial would be 0). You should press 1 if you want to take the action and 2 if you do not.

The amount of points you can gain or lose as well as the probability of success will change on each trial. You can think of this as depicting real-life situations in which the probability of success and the possible outcomes of your actions will be different depending on changing factors such as the time and resources you would have needed to perform the action.

Critically, in most real situations, you never have full and complete information about the chances that your action will succeed. Thus here as well, on some trials you will not have the full information concerning the amounts of green and red slices in the circle. In these trials the circle will be partially obscured by a gray cover which will not allow you to see that part of the circle (see example below).



You have an opportunity to acquire food by trying to fish in a small close river.

Observable red area = 30 %
Observable green area = 20 %
Observable gray area = 50%

Possible gain: +50
Possible loss: -25

Select: '1' to fish

Select: '2' not to fish

In this case 50% of the circle is under the gray cover and is not known to you. Under the gray cover can be any mix of red or green sections.

Please note that this summarizes what you can see out of the total circle.

In this example, the total circle includes at least 30% red sections, at least 20% green sections and 50% of the circle is unknown to you.

The amount of information available to you (i.e. the size and position of the gray cover) will be different for each decision.

VII. Appendix S2. Example instructions for the Delgation task.

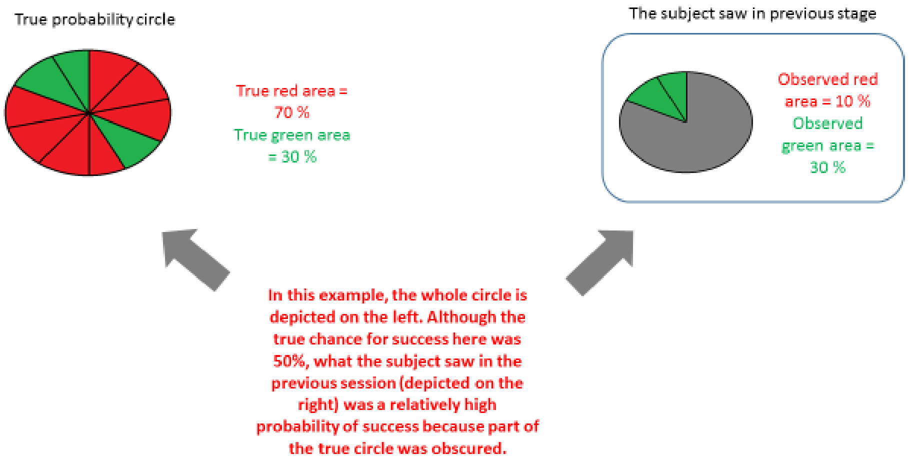
Welcome to the Econ-Lab

You are now participating in an experiment at the Department of Economics, University of Zurich. Please read the following instructions carefully for at least 15 minutes, your full understanding of the instructions will allow you to achieve the best possible outcomes in terms of CHF earned for you and your group members. Please take the time to understand the instructions and how the outcome of the payment is determined. Also note that in the end of these instruction pages you will need to answer a short quiz to assess you understood the instructions.

Today you will perform the same choices you made in the previous stage. As before, on each trial you can select whether to perform an action or not. Each action has a probability of success as indicated by the amount of green (success) and red (failure) in the circle. Each action also has a possible gain if it succeeds and a possible loss if it fails. If you decide not to take the action you will not lose or win anything for that trial (outcome for that trial = 0 CHF). As before, a gray cover in changing sizes will prevent you from seeing the entire information in the circle.

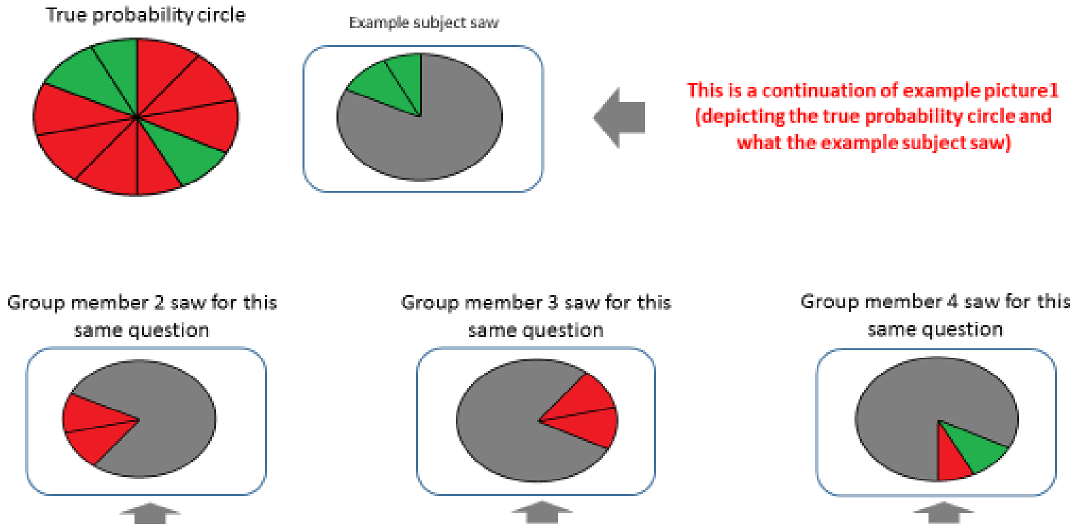
However, in today's session you will be able to defer to the option preferred by the majority of your group. In the previous stage, all your group members saw the <u>same questions with the same size of gray space</u>. However, the position of this gray space was different for each individual (see example below). Since each member of the group saw a different piece of the picture, on average the whole group can have more information on the true amounts of red and green slices in the circle then any individual member (see example below).

Example picture1: the true probability vs. what one example subject could see:



In this example, the whole circle is depicted on the left. Although the true chance for success here was 50%, what the subject saw in the previous session (depicted on the right) was a relatively high probability of success because part of the true circle was obscured.

Example picture2: the same true probability vs. what other group members could see:


True probability circle
Example subject saw
Group member 2 saw for this same question
Group member 3 saw for this same question
Group member 4 saw for this same question

This is a continuation of example picture1 (depicting the true probability circle and what the example subject saw)

Here however you can also see what each of the other group members saw. Each member saw the exact same question and probability circle but each had a different part of the circle obscured. As you can see, the majority of the remaining group members (members 2 and 3) may have made a different decision than the example subject in this case.

As you can see from the example above, although each individual is exposed to the same amount of information, the group as a whole had more information than most individuals. This phenomenon is often referred to as “wisdom of the crowds” and forms the base for popular websites such as TripAdvisor.

In fact, as you can intuitively imagine, the wisdom of the crowds’ phenomenon provides a mathematically quantifiable informational advantage, mostly when the individual has very partial information (i.e. when there is a large gray area). When the obscured space is large, each individual will be making a bet based on very limited information and the group average will have a large informational advantage over the individual. As the obscured space gets smaller, mathematically, the group will have a smaller informational advantage over the individual. Intuitively this is because when the information each individual sees is similar the advantage of getting multiple independent perspectives is no longer present.

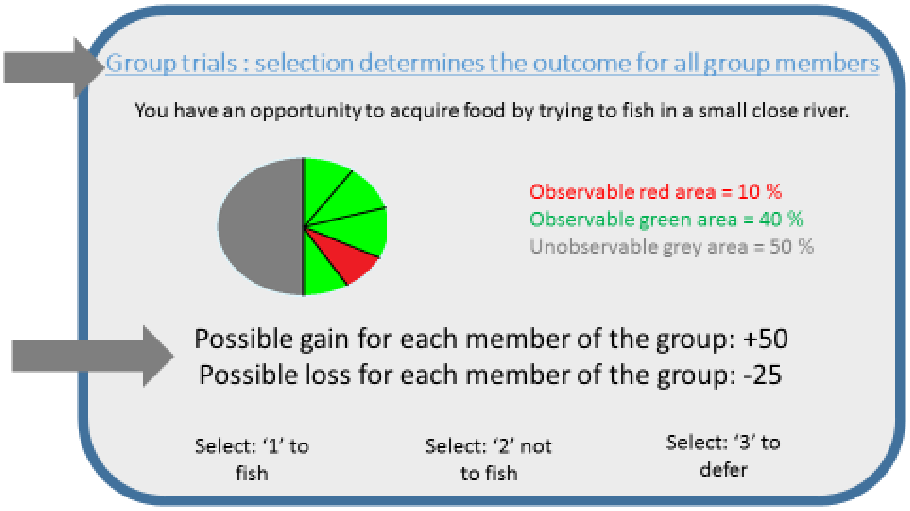
In today’s session, you could decide on each trial to either take or not take an action. However, you can also let the group make this decision (*Deferring*-see example picture below). In this case the option preferred by most of your group members (not including yourself) on this specific question in the previous session will be selected for the whole group (*including yourself*). So for the example above, if the example subject decided to defer on this trial, this means s/he is allowing the decision of the majority (i.e. members 2,3,4, not including the example subject him/herself) in the previous stage to determine his/her action now.

Please remember that although the group as a whole may have more information than each individual, each group member has their own personal preferences (for example, in terms of how much risk they are willing to take) and thus your group members may make very different decisions than you. For example, previous experiments have demonstrated that some participants are willing to risk large amounts in order to obtain large gains, whereas other participants prefer in such cases to take a safe option.

Critically, in today's session there will be two trial types (Group, Private) that are explained below. Please note that as in the previous session, in order for you to pay full attention to the question you will not be able to answer in the first 4 seconds.

Group action and consequence trials (Blue background trials).

If you see a blue background screen, you are in a "Group" trial. In these trials, your whole group is dependent on you as their leader. You will all face the obstacle as a group and your action as the leader of the group will determine the actual money earned by **each and every one of your group members**. That is, the money resulting from your choice will be given to each of your group members (including yourself). See example picture below.



Group trials : selection determines the outcome for all group members

You have an opportunity to acquire food by trying to fish in a small close river.

Observable red area = 10 %
Observable green area = 40 %
Unobservable grey area = 50 %

Possible gain for each member of the group: +50
Possible loss for each member of the group: -25

Select: '1' to fish

Select: '2' not to fish

Select: '3' to defer

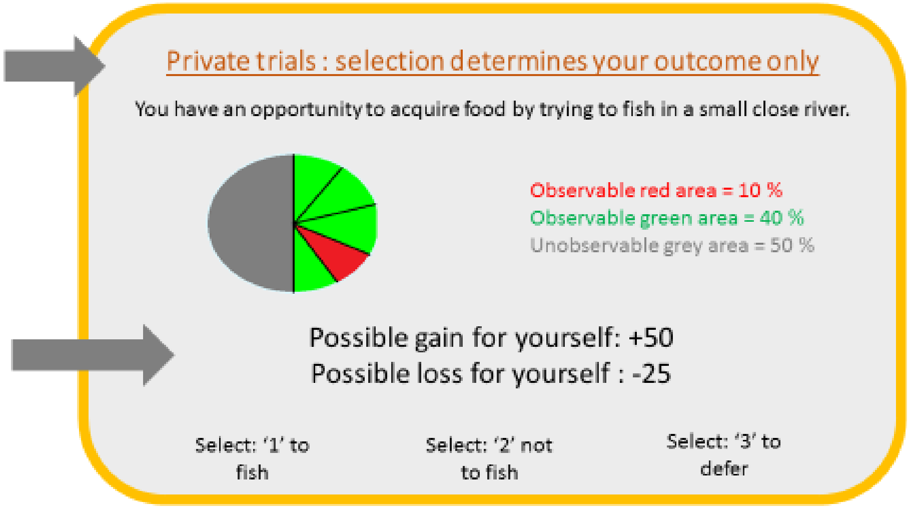
The background color and the title indicate you are in a group trial. Remember, for these trials you are the leader of your whole group and your choice will directly affect the payment of each and every group member.

Any gain or loss resulting from your decision will be directly added or removed from the payment of each of your group members (including yourself)

Please select the 1 key on your keyboard to take the action, the 2 key not to take the action and the 3 key for deferring. As mentioned before deferring means that the selection for this trial (either to act or not to act) will be determined by the answers of the majority of your fellow group members (not including yourself) in response to this exact question in the previous stage. For group (blue background) trials, the outcome, no matter if you decided to lead or defer, will be added/subtracted from the earnings of each group member including yourself.

Private action and consequence trials (Yellow background trials).

If you see a yellow background screen, you are in a **'Private'** trial. In these trials, the challenges will be identical to the Group trials, however, your selections will have absolutely no consequence for your group members, and will affect yourself only. These trials are paid out separately and do not depend on the group action trials. As before, you could defer to the group majority if you chose to rely on the opinions of your fellow group members. In this case the majority opinion will determine the action selected, but the outcome will be for yourself alone (see example picture below). Thus for Private trials the outcome, no matter if you decided to lead or defer, will be added/subtracted from your earnings only.



Private trials : selection determines your outcome only

You have an opportunity to acquire food by trying to fish in a small close river.

Observable red area = 10 %
Observable green area = 40 %
Unobservable grey area = 50 %

Possible gain for yourself: +50
Possible loss for yourself : -25

Select: '1' to fish

Select: '2' not to fish

Select: '3' to defer

The background color and the title indicate you are in a private trial. Remember, for these trials your decisions will affect yourself only. No group member will be influenced by your actions.

Any gain or loss resulting from your decision will be added or removed from your payment only.

Payment structure:

All your fellow group members will be doing this task, the results of each group member (**for the Group trials only**) will be added or subtracted from the outcome of each other group member. Thus, your performance (on Group trials only) will significantly affect not only your outcome but the outcome of each of your group members.

Your final payments will be the sum of the total performance of all your group members (including yourself) on Group trials, with the addition of your personal payment from Private trials. As you can see from this payment structure, the amount of money you can personally win/lose from your Group or Private trials is the same (this is under your control). In addition, your payment will also depend on the performance of your group members in their group trials but this is not under your control.

Each point is always worth 0.4 CHF. As in the previous stage 5 trials will be randomly selected from the group trials and 5 trials from the private trials. The abovementioned outcomes from this stage will be calculated based on the outcomes from these trials only. Thus, it is critical that you pay attention to each trial and make the selection you prefer on each trial.

General information about the next stage:

At the end of this stage, all the members of your group will meet together with the experimenter. The amount of money each individual accumulated for the group will be announced. At this stage the payment for all previous stages will be given.