The Cost of Appearing Suspicious? Information Gathering Costs in Trust Decisions

Ma, I.[1,2], Sanfey, A. G.[1], Ma, W.J.[2]
[1]Donders Institute, Radboud University, The Netherlands
[2]New York University, New York, USA

Abstract

Trust decisions are inherently uncertain, as people usually have incomplete information about the trustworthiness of the other prior to their decision. Therefore, it is typically beneficial to gather information about another's past behaviour before deciding whether or not to trust them. However, gathering elaborate information about someone's history could convey distrust in them, and thereby impact their behaviour. Here, we examined how people balance the benefit of information acquisition against the cost – social, monetary, and both – of gathering this information. To do this, we gave participants the opportunity to sequentially sample information about that trustee's reciprocation history before they decided whether or not to trust. A financial cost was added by making participants sometimes pay for this information, and a social cost was induced by telling participants that the trustee would sometimes learn about their sampling behaviour. As expected, people sampled less when there was a monetary cost to acquiring prior information. More surprisingly, when sampling was free, people sampled less when the trustee would later be informed of their acquired information quantity, suggesting a social cost of appearing distrustful. Within each cost condition, people sampled more when the reciprocation history was less conclusive. To better understand the computational strategies employed, we compared a Bayesian Optimal model, a Bayesian heuristic model, and a simple Drift Diffusion model. We find that with the Bayesian heuristic model fitted the data best, which sheds light on how people use their posterior beliefs over trustworthiness in decisions to sample and trust. This study opens the door to broader applications of the tools and models of information sampling to understand social decision-making.

Keywords: Trust, social decision-making, uncertainty, information sampling, Bayesian computational modelling

The processes of trust and reciprocity are essential for both establishing and maintaining beneficial human cooperative interactions[1]. When trust is invested, and then reciprocated in turn, the investor (who makes the decision to trust) and the trustee (in whom trust is placed) typically emerge better off from the exchange. However, a trustee can often maximize their own profit by not reciprocating the trust placed in them. Therefore, perhaps the key factor in a successful cooperative interchange is the investor's belief about whether or not the trustee can in fact be trusted. Understanding how an investor both gathers and uses information to make this choice is a fundamental question in the study of social interaction.

Obviously, the ever-present possibility that one's trust might be violated confers inherent uncertainty on the decision of whether or not to trust someone. Some individuals are more trustworthy than others, and repeatedly interacting with the same person does not necessarily result in the same outcome[1,2]. In addition, we often only have limited information about someone's level of trustworthiness. Therefore, decisions that involve trust are both risky and uncertain: there is variability in the outcome given a specific degree of trustworthiness of the other person (risk), and there is incomplete knowledge about this trustworthiness (uncertainty, sometimes also referred to as ambiguity). Due to this uncertainty, it is beneficial to acquire information to determine whether someone can be trusted, which can be done by gathering information about the trustee's past behaviour[3].

Information gathering of this type is usually costly in terms of time, effort, or money. Prior experiments examining information gathering have typically assessed non-social decision contexts, for example when assessing which risky gamble may pay off at a greater rate. These studies demonstrate that, not surprisingly, the imposition of costs cause people to acquire less information and subsequently make less optimal decisions[4,5]. Therefore, our first research question is: Do people sample information to decide to trust another or not and are they willing to pay financial costs to acquire this information? Thereby, we extend the information sampling literature into the social domain.

The aforementioned question is complicated however by the fact that gathering information about other people can involve additional social costs. Acquiring information about others requires that you consider the consequences of your inquiries on the other person's impression of you[6,7]. For example, too many detailed questions about another's history of trustworthiness could be seen as offensive and thereby in turn affect the other persons' behaviour.

If someone continually checks up on our reliability, it may make us less likely to behave pro-socially with that person in the future, which leads to our second research question. Do people take into account how our information gathering may be interpreted by their interaction partner?

Previous studies have shown that people care about the intentions of others in cooperative decisions[8,9]. Information gathering can sometimes reveal the investors' covert intentions that govern the cooperative decision-making process. For example, it has been proposed that some individuals care about whether their cooperation partner gathers information about their own profit if they were to defect, before making a cooperative decision that is mutually beneficial[10-12]. Under that assumption, simulations of cooperative games show that blind cooperation evolves to become a beneficial strategy[10-12]. For example, when individuals compete to become cooperative partners in heterogeneous populations with both reliable and unreliable individuals, those individuals who blindly cooperate should be preferred[11]. The assumption of these simulations is corroborated by behavioural experiments showing that people who gather information about their own profit are perceived as more selfish, and indeed generally also behave more selfishly[7]. This converging evidence suggests that cooperative decision-making incorporates a social cost to information gathering when assessing the potential payoff structure. Here, we extend this concept of a social sampling cost from structural information to interpersonal information, that is, sampling information about the trustworthiness of others.

Our third goal is to understand the computational mechanisms by which people may sample to decide whether someone can be trusted. We therefore compared three computational models: a Bayesian Optimal model, a Bayesian heuristic model, and a Drift Diffusion model (Methods).

To answer these questions, we designed a novel version of a single-shot trust game, the Information Sampling Trust game (IST). This allowed us to investigate information sampling in trust decisions and, within-subjects, compare between overt (trustee is informed of sampling) and covert (trustee is not informed of sampling) information acquisition. Additionally, participants either could sample for free, or had to pay per sample. Participants completed the IST in the role of investor, and were given the opportunity to sequentially gather information about a trustee's previous reciprocation history before deciding whether or not to trust (Figure 1). The sequentially revealed trustee's reciprocation history consisted of independent draws from a Bernoulli distribution with parameter equal to the trustee's reciprocation probability r. On each

trial, r was drawn pseudo-randomly from six possible values (r = 0, 0.2, 0.4, 0.6, 0.8, 1). After the investor decided to finish sampling, they chose whether or not to trust (invest). The experiment consisted of four conditions: a monetary cost of sampling (either costly or cost-free), crossed with social context (trustee informed or not informed of how many boxes the investor opened). Of the trials on which the investor decided to trust, three trials were randomly drawn and the average of the outcomes on these trials was paid.

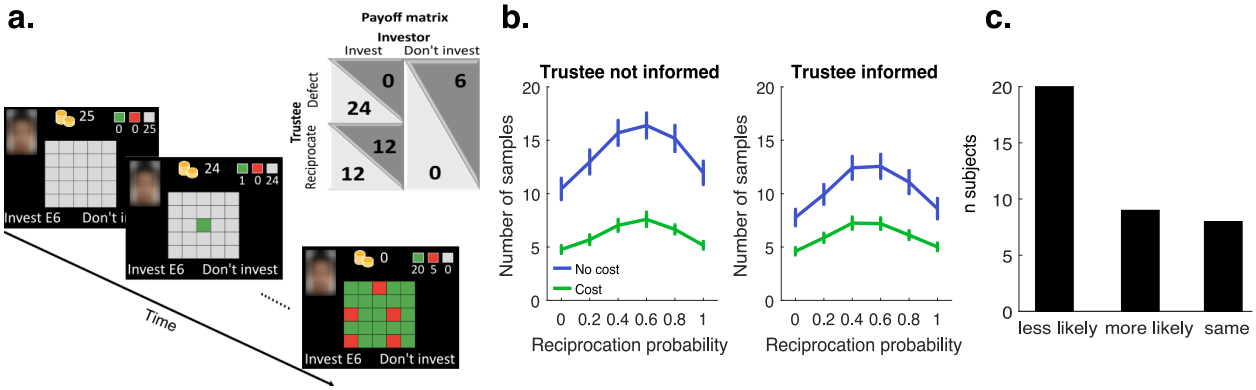

Figure 1. Task and data. A. Trial sequence in the Information Sampling Trust game (IST) with the payoff matrix. The participant played the task in the role of investor. Similar to the original Trust Game[1], the investor was endowed with €6 at the start of each trial. This endowment could either be entrusted to (invested in) the trustee or kept without investing. The investor was told that if the endowment was entrusted to the trustee, then the trustee's acquired amount would be multiplied by 4. The trustee then subsequently decided to (a) reciprocate trust by returning 50% of the multiplied endowment to the investor, or (b) defect/betray, which meant keeping all the money, so that the investor ended with 0. Crucially, before the investor made a trust decision, they could sequentially sample a trustee's reciprocation history with other investors. On each trial, information could be sampled up to 25 times. The colour of the opened box indicated the trustee's past decision. Green = reciprocated trust, red = did not reciprocate trust, grey = not sampled. The investor could stop sampling at any time, and then had to decide to either trust (invest €6) or not trust (keep €6). Investment outcomes were not shown during the task. In the monetary cost conditions, 5 eurocents were deducted for each decision to sample (from a separate sampling budget of 125 eurocents per trial). After each investment decision, a new trial began with a new trustee. Each trustee was encountered only once. B. Data demonstrates that people sampled less when the trustee was informed of how much information was sampled. Mean and SEM of drawn samples for each generative reciprocation probability per cost condition: green = monetary cost, blue = free. Left panel = trustee not informed, right panel = trustee informed. C. The most common belief amongst participants was that the more information they overtly sampled, the less likely reciprocation would become. Participants could choose from one of 3 categories: overtly sampling more information would make reciprocation less likely, more likely, or stay the same.

To assess the effect of cost condition and sample outcomes on the decision to sample or stop, we conducted a mixed-effects model (SI). There was a significant interaction between monetary and social cost manipulations (coefficients mean±SEM; -0.10±0.01, $p < 0.001$); participants sampled less if the trustee would be informed of the sampling when sampling was monetarily cost-free (-0.45±0.04, $p < 0.001$) but not when samples were monetarily costly (0.04±0.04, $p = 0.37$). Participants also sampled less when sampling was monetarily costly compared to cost-free, both when the trustee was informed (beta = -0.64±0.04, $p < 0.001$), or not informed (-1.06±0.04, $p < .001$). As expected, the distance between reciprocation probability r and 0.5 (maximum risk) also affected sampling (0.41±0.01, $p < 0.001$). Specifically, information sampling increased when the samples were drawn from an r closer to 0.5, and the acquired information was thus relatively inconclusive. This effect of r on sampling was not completely symmetric with respect to $r = 0.5$ (Figure 3). Specifically, Bonferroni-corrected pairwise t-tests revealed that people sampled more when $r = 0.8$ than when $r = 0.2$ ($t(40) = -2.05$, $p < 0.001$, $d = 0.59$) but not for $r = 0.4$ compared to $r = 0.6$ ($t(40) = -1.44$, $p = 0.010$, $d = 0.42$) or for $r = 0$ compared with $r = 1$ ($t(40) = -0.88$, $p = 0.085$, $d = 0.28$).

We then used a separate logistic regression to test whether the decision to trust itself was completed was predicted by r and the cost conditions. This logistic regression returned a coefficient 9.12±0.18 ($p < 0.001$) for r, indicating that the probability of trusting increased with a higher r. This confirms that the acquired information was actually used in the final trust decision. As expected, the two cost manipulations (monetary and social) and their interaction were not significant predictors of the decision to trust (monetary cost: -0.023±0.095, $p = 0.81$; social cost: 0.011±0.096, $p = 0.91$; interaction between monetary and social cost: 0.113±0.135, $p > 0.403$). After task completion, participants were asked to indicate whether they believed that sampling made reciprocation more likely, less likely, or the same when the trustee was informed of the information gathering. Believing that overt information sampling would make reciprocation less likely was the most commonly-reported answer ($\chi^2(2, n = 37) = 19.28$, $p < 0.001$, Figure 1c).

For the Bayesian Optimal model and the Uncertainty model, we consider the possibility that people use their uncertainty over trustworthiness in their decisions to sample. Under these models, people maintain a posterior belief distribution over another's trustworthiness (Methods).

Earlier sample outcomes then update the posterior belief more strongly than later ones, as uncertainty over trustworthiness is highest when information is sparse. As more information is sampled, uncertainty decreases and new samples will not change the posterior much. The posterior distribution can be used in a decision to gather more information in two distinct ways: Optimal or heuristic.

The *Optimal model* (Methods) uses the posterior to calculate which action has the highest expected utility: sampling, or stopping (to trust, or not trust). We derived the expected value of a state-action pair using the Bellman equations and dynamic programming[13], taking into account the immediate subjective cost of sampling, *c*; we allowed this cost to vary between the four cost conditions.

The implementation of this optimal strategy is likely computationally expensive and the brain might instead use a "good-enough", simpler strategy[14,15]. We tested a suboptimal Bayesian strategy in the form of the *Uncertainty Model* (Methods). Here, the agent keeps track of their uncertainty over trustworthiness as measured by the width of the posterior belief distribution. The more information is sampled, the narrower the posterior becomes. The agent stops sampling when uncertainty drops below a criterion *c*; we allowed this criterion to vary between cost conditions.

An alternative to the Bayesian view is that people do not use posterior belief distributions. Instead, they may maintain criteria for when they view people as trustworthy or untrustworthy. This is captured by a *Drift Diffusion model* (DDM)[16] (Methods). Here, the agent samples information to test whether the trustee meets their criterion for either trustworthiness or untrustworthiness. This requires keeping track of the sample outcomes in favour of trusting and not trusting. The idea is that this difference can act as an approximation of certainty. The decision to stop sampling information is then determined by whether their difference is sufficiently large, i.e., when the difference reaches a bound *b*. We allowed the bound to vary between cost conditions. Compared to the Bayesian view, this model predicts that people are not sensitive to the absolute amount of gathered information, only to the difference between positive and negative outcomes.

We first tested several versions of each of the three models (summarized in figure 2). For the Optimal model, the best fitting version included subjective prior beliefs[17] and risk aversion (often referred to as betrayal aversion in the context of trust)[18]. The best Uncertainty model

version included a subjective prior. However, for the DDM, using neither separate nor collapsing bounds convincingly improved the model fit (for AIC and BIC results see SI).

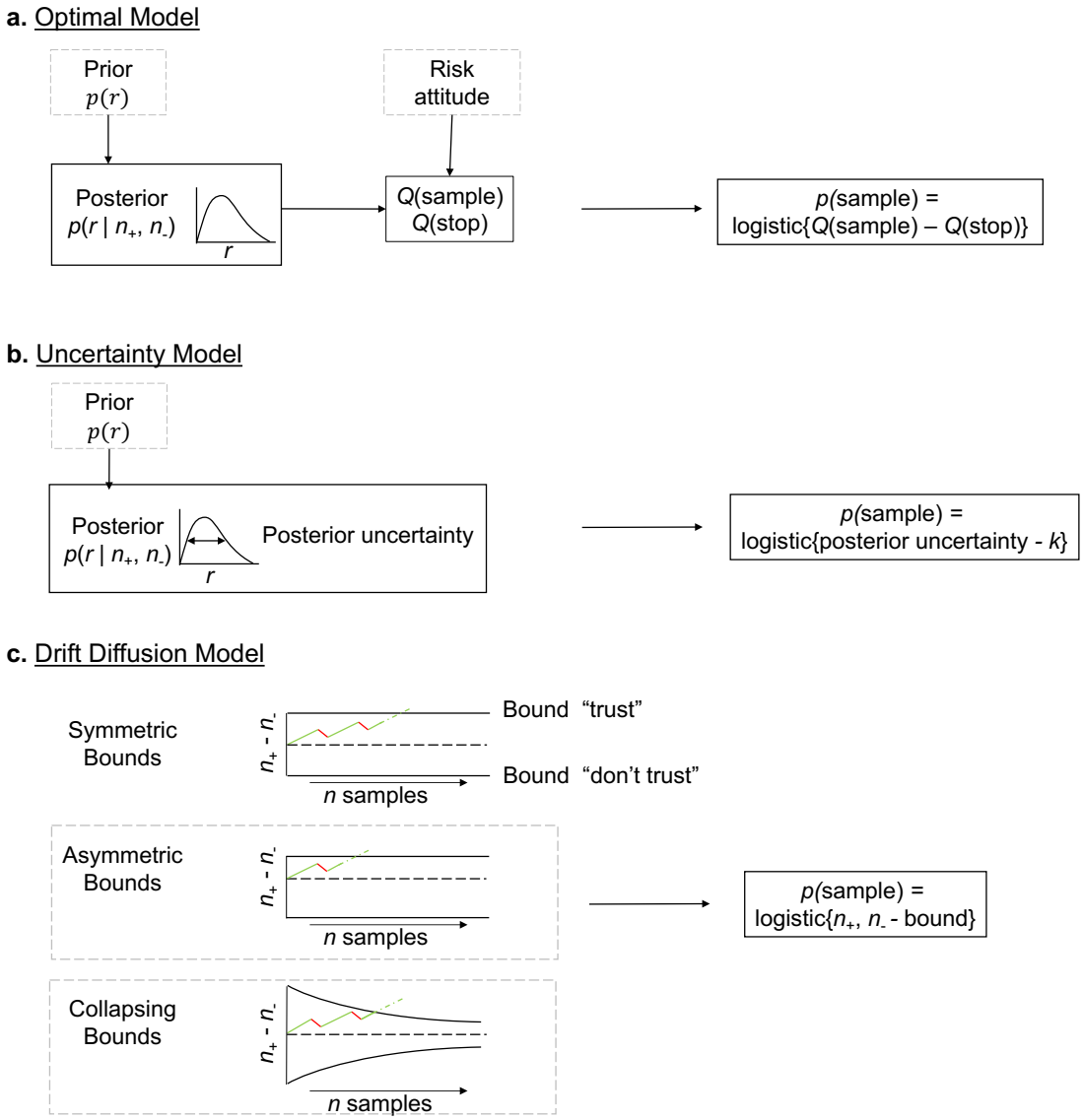

Figure 2. Overview of the computational models. The number of opened boxes with a positive outcome n_+ reflects the number of opened boxes with positive outcome, n_- is the number of opened boxes with a negative outcome. A. The Optimal model, which uses a Bayesian belief distribution over the probability of reciprocation. The prior was uniform in the basic model variant. We considered model variants with subjective prior beliefs and/or potential risk aversion (grey boxes). B. Uncertainty model. Information is sampled until the uncertainty drops below a threshold. In the basic model variant, the prior was uniform. We considered a model variant with subjective prior beliefs (grey dashed boxes). C. The Drift Diffusion model. The agent accumulates evidence, computing the difference in positive (green) and negative (red) samples until a bound is reached. The green dotted line is meant to indicate that this bound is not a "hard"

bound at which people stop sampling for sure. Instead, we allow for decision noise, which means that the probability of stopping increases as the bound is approached and exceeded. In the basic model variant, the bounds were symmetric and fixed. We examined model variants with asymmetric or collapsing bounds. In all models, we allowed for decision noise.

We then compared the best performing version of each model and showed that the Uncertainty model performed better than the Optimal model (AIC = 2082 95% CI [635, 3597], BIC = 2504 95% CI [1065, 4023]). Moreover, the Optimal model fitted better than both the DDM (AIC = 3974 bootstrapped 95% CI [2767, 5239], BIC = 3130 95% CI [1920, 4397]). The Uncertainty model also fitted better than the DDM (AIC = 6056 95% CI [5117, 7055], BIC = 5634 95% CI [4689, 6639]). Thus, out of the 3 models, the Uncertainty model fitted best (figure 3a, b) and performed well in all cost conditions (figure 3c).

Moreover, the Uncertainty model's criterion estimates confirmed that people were more tolerant to uncertainty when the trustee was informed of the sampling, and when sampling was monetarily costly. Specifically, the criterion estimates differed between conditions (Wilcoxon Signed-Ranks test all p <.005, see SI), except for the comparison between informed and not informed trustees when samples were monetarily costly. On average, people maintained priors means that were slightly negatively biased (median prior estimate = 0.475, Wilcoxon signed-rank returned: Z = -2.738, p = 0.006). This bias allowed the model to account for the empirical finding that people gathered more information when the generative r was 0.8 than when r was 0.2. In sum, model comparisons demonstrated that people do not sample in accordance with an optimal strategy, but rather sample until uncertainty drops below a subjective criterion. This criterion depended on the posterior probability distribution over reciprocation probabilities. More specifically, we excluded a family of simple DDM's, including a DDM with assymetric bounds and a DDM with collapsing bounds, as a human strategy for trustworthiness sampling in this experimental context.

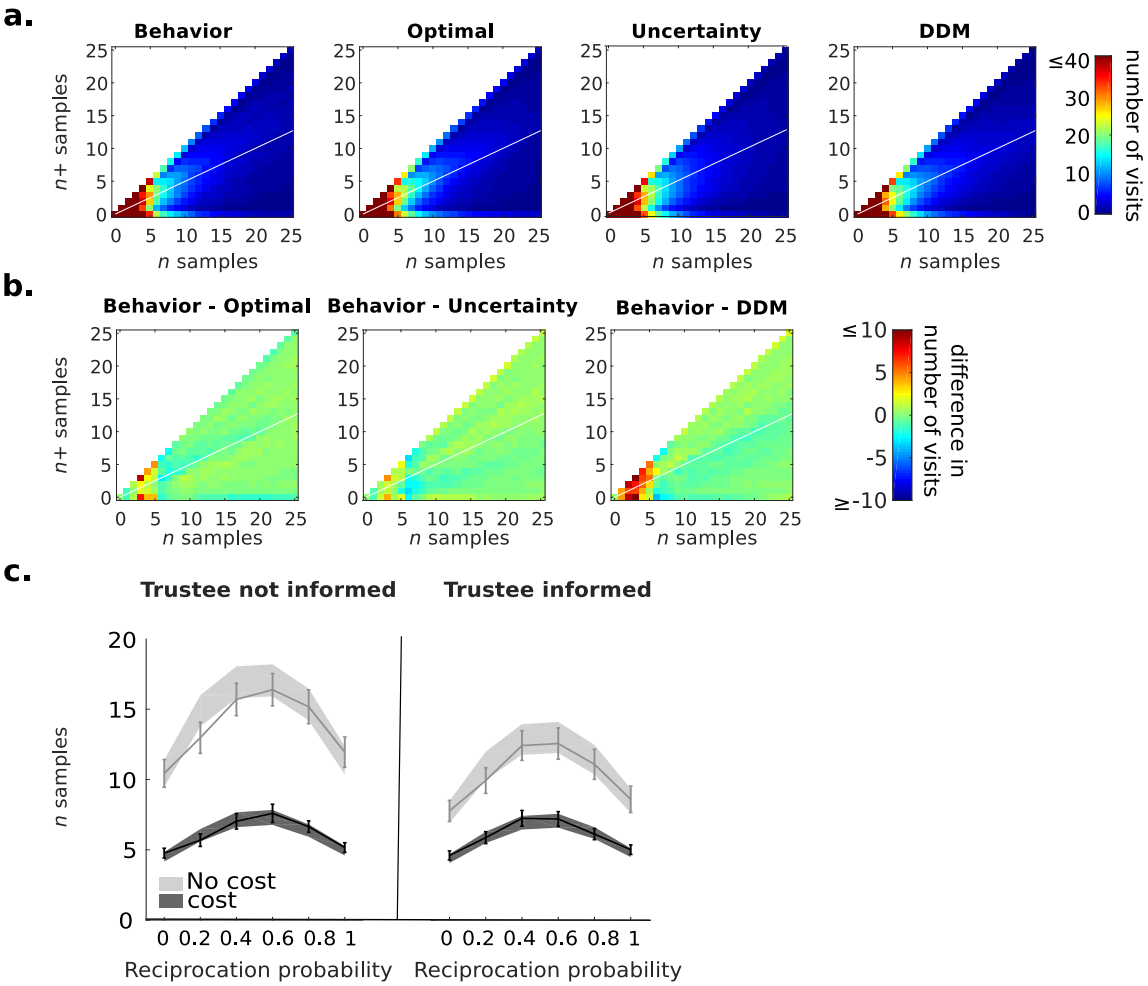

Figure 3. Model fits for the sampling decisions. A. Number of times that each possible state was visited. Plotted for the raw data and synthetic data as generated by the models using estimated parameters, averaged over participants and cost conditions. The full state space is determined by every possible combination between the number of samples (*n* samples) and number of positive samples (n_+ samples). Note that reciprocation probabilities of 0 and 1 follow one specific path in the state space, whereas other values of *r* "spread out" in the middle due to stochasticity in the outcome of a sample. Therefore, trials in which participants reached the states along the upper edge ($r = 1$) and bottom edge ($r = 0$) paths occur in high frequency, even though participants sampled more when *r* was closer to 0.5 (as depicted in plot a; white line through the middle is included as a reference point). B. Difference between the behaviour and data synthesized by the models in the number of visits. This highlights that the DDM fits less well than the Optimal and Uncertainty models. C. Number of samples (*n* samples) per cost condition and the Uncertainty model fit. The raw data is represented by lines (mean) and error bars (SEM) across all subjects. Shaded areas indicate SEM of the model fits.

Next, we considered individual differences, as the population may well be heterogeneous, in the sense that different individuals follow different models. We therefore used Bayesian

Model Selection[19], which returned the following expected frequencies of the models in the population: 0.40 for the Optimal model, 0.59 for the Uncertainty model and 0.01 for the DDM. This suggests that the majority of individuals employ a posterior distribution over *r*, instead of using a difference of sample outcomes as their decision variable. The exceedance probabilities, which reflect the probability that a given model is more likely than any other model, returned: 0.12 for the Optimal model, 0.88 for the Uncertainty model and 0 for the DDM. This is strong evidence in favor of the Uncertainty model as the most likely model in the population.

We further examined whether the two winning models could also predict trust decisions (after sampling has concluded) by using the parameter estimates. We therefore extracted the subjective expected value of trusting at the time of stopping for the Optimal model and Uncertainty model and for each model separately, we fitted the subjective expected values to the trust decisions, allowing for bias and decision noise temperature (equation 9). The fits were not significantly different (95% CI of the summed difference in log likelihoods [-54.66, 46.44]). Both models resulted in a significant predictor of the probability of trusting (Uncertainty model $p < 0.001$, Nagelkerke pseudo-$R^2 = 0.882$, Figure 4a; Optimal model $p < 0.001$, pseudo-$R^2 = 0.832$, Figure 4b; DDM, $p < 0.001$, pseudo-$R^2 = 0.733$).

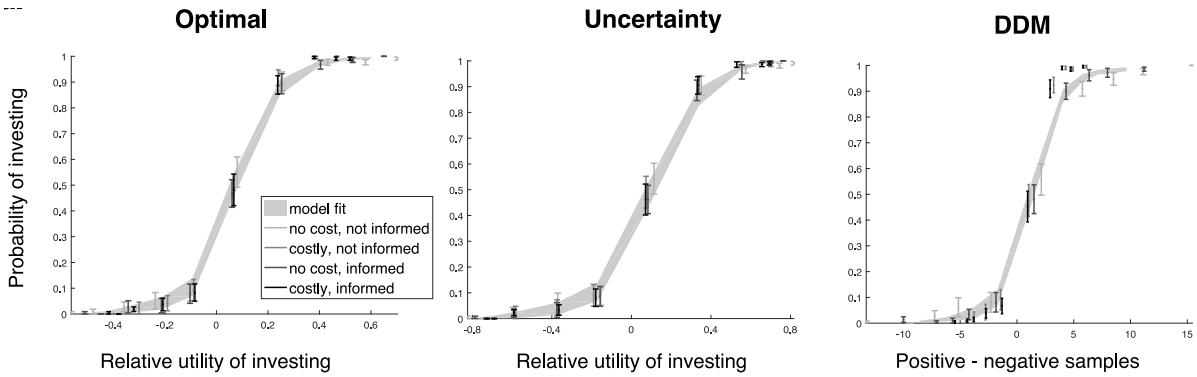

Figure 4. Model fits to the trust decisions for each model using their decision variables. For the Optimal and Uncertainty models the decisions to trust are plotted as a function of expected utility of trusting derived from the models. For the DDM, the decision variable is the difference between positive and negative samples. Error bars represent the data (mean) ± SEM across all subjects. Shaded area is the model fit ± SEM.

To test the robustness of our findings under variations of the distribution of the reciprocation probability, we conducted a second, independent study ($n = 75$; see SI) with biased generative distributions of *r*. The experimental procedure was identical to study 1, apart from the

fact that participants sampled over either positively biased (r = .2, .4, .6, .8, and 1), or negatively biased (r = 0, .2, .4, .6, and .8) generative distributions. Our findings of social and monetary costs replicated (SI). Within-model comparisons replicated for the Optimal and Uncertainty models. However, for the DDM, asymmetric bounds improved the model fit (SI). Between-model comparisons did not distinguish well between the Optimal and Uncertainty models (AIC = -238 95% CI [-2365, 1731], BIC = 626 95% CI [-1498, 2600]) but again the Optimal model performed better than the DDM (AIC = 6761 95% CI [5319, 8373], BIC = 2873 95% CI [1417, 4505]), as did the Uncertainty model (AIC = 3500 95% CI [2006, 4908], BIC = 3500 95% CI [2024, 4893]). The following expected frequencies of the models in the population: 0.40 for the Optimal model, 0.49 for the Uncertainty model and 0.11 for the DDM. This again shows that using the difference as a proxy for uncertainty is uncommon. The exceedance probabilities returned 0.20 for the Optimal model, 0.80 for the Uncertainty model and 0 for the DDM. This shows that the Uncertainty model is still the more likely than the other two models. In contrast to study 1, the Uncertainty model better predicted the trust decisions after sampling concluded (Nagelkerke pseudo R-squared = 0.64, $p < .001$) compared with the Optimal model (Nagelkerke pseudo R-squared = 0.15, $p < .001$). These findings further substantiate the suggestion that the computational mechanism described by the Uncertainty model may be closest to the human strategy.

Across two studies, we demonstrated a robust social cost effect: people discounted the value of information when the acquisition of that information was overt to the trustee. Crucially, we did not instruct our participants about whether, or how, information sampling might affect reciprocation probabilities. Further, sequential information acquisition and subsequent trust decisions were better explained by a Bayesian heuristic model, the Uncertainty model, than by a Bayesian Optimal or a Drift Diffusion model (DDM). In a replication study, the Uncertainty model and the Optimal model both fitted the sampling decisions better than the DDM, but were not clearly distinguishable from each other. The best fitting models suggest that people build a posterior belief distribution over trustworthiness, and then use these beliefs for subsequent decisions, including whether to gather more information.

The social cost finding is intriguing, as it shows that when people gather information about another person, they factor in that these inquiries may change that person's behaviour

towards them. In this regard, learning about others diverges in important ways from information acquisition regarding the statistics of a non-social environment. The former requires thinking about the effects that one's inquiries may have on others. This reasoning about how other peoples' behaviour may change as a function of your own actions appears therefore to require theory of mind[20]. This is further supported by participants' self-reports, which indicated that they believed excessive information gathering would decrease the trustee's probability of reciprocation.

An interesting question in this regard is whether this belief is in fact adaptive, even though it counterintuitively leads to decisions being made using *less* information. Previous work might shed some light on this question. Looking for better options while in a potentially reciprocal partnership will eventually decrease the probability of the continuation of that partnership[10,12]. People then often decide not to overtly look for better options in order to signal to their partner that their cooperative behaviour is unconditional, i.e. independent of the value of the "temptation" to defect[6]. Our findings extend this work in important directions, as in our task information was sampled about the trustee's decision history itself, instead of about other potentially tempting options. This shows that people reduce their quantity of information sampling to clearly signal that they are not *suspicious* of the trustee. This social information gathering cost thereby demonstrates that the appearance of trust in another contributes to the decisions to acquire information for the purpose of building trust.

As mentioned, one potentially negative consequence of this social cost is that the subsequent trust decisions are more poorly informed, as they are made with fewer items of information. Many studies have revealed biases that arise from decisions based on limited information. In particular, negative initial impressions often result in strong learning biases as they discourage future interactions and thereby obstruct learning via avoidance. Initial positive impressions typically have the opposite effect, by encouraging future interactions. Therefore, positive impressions of another person are generally more susceptible to updates from experience by successive interactions with that same person[4]. Other studies have shown that prior beliefs about someone's trustworthiness based on facial judgement of trustworthiness, attractiveness, narratives about moral character, and previous social experiences in unrelated settings, can hinder learning about trustworthiness[17,21-23]. Given that social costs actually reduce information search, decisions might become more susceptible to these biases.

In addition to, and independent of, the social cost effect, participants were clearly sensitive to uncertainty. People gathered most information when r was close to 0.5, and least information when r was deterministic. While this is qualitatively optimal, participants also showed an interesting valence-dependent asymmetry in information gathering. Specifically, when outcomes were stochastic, people sampled more when outcomes were mostly positive and less when outcomes were mostly negative. This asymmetry in sampling indicates that beliefs about trustworthiness are more likely to be updated than beliefs about untrustworthiness. At first glance, this asymmetry in sampling may seem consistent with well-established findings in impression formation and person perception, in which negative behaviour is generally viewed as more diagnostic of morality than positive behaviour[24-27]. However, those studies predict that negative evidence is especially more heavily weighed when behaviour is extreme, which in our study would be the case when trust was never reciprocated, i.e. $r = 0$. In contrast, our study revealed that people showed an asymmetry in sampling when sample outcomes were stochastic instead of deterministic. This suggests that a negativity bias in trustworthiness information search interacts with information consistency, such that people are suspicious of others who are mostly trustworthy but not consistently so.

Despite our rigorous model comparisons and robust findings, claims about deviations from optimality are limited to some extent by our use of binary sampling costs. Parametrically varying social costs in future studies may require informing the trustee about sampling behaviour in a stochastic manner. A high probability of informing the trustee should then induce a high social cost of information sampling. Another potential extension could involve directly probing the trustee's behaviour. Understanding whether, and how, trustees may change their behaviour when investors gather information is important in understanding whether the social cost we observe here is truly adaptive. These approaches may also shed more light on why people may follow the Uncertainty model rather than the Optimal model. One possibility is that people choose a heuristic strategy that closely mimics the optimal strategy but that is computationally less expensive, because there is no need to look ahead to the end of the trial. Making this speculation concrete would require formalizing the computational costs or constraints and calculating optimal strategies under such costs or constraints, as in bounded rationality[28,29] or resource rationality[30,31].

We believe that the approach we have utilized here has the potential to open many interesting doors in exploring the nature of social interactive decision-making. For example, our findings provide a benchmark to uncover social value computation aberrations in psychiatric disorders. Some psychiatric disorders can be characterized by biases in information sampling, such as insufficient information gathering in addiction[5], asymmetric weighting of negative evidence in depression [32], and impaired information sampling costs signals associated with compulsivity [33,34]. Importantly, specific impairments in the ability to model the moral character of others or respond to social signals are central to a range of psychiatric disorders, including borderline personality disorder and autism spectrum disorder[35,36]. For example, explicitly modelling the another person's state of mind, including dynamic changes in their intentions, can identify maladaptive social behaviours[37]. Our study further adds to the mathematical formulations of these interpersonal interactions by applying computational modelling approaches from information sampling to social dilemmas.

METHODS

Study 1

Participants

A total of 40 participants (13 males) completed the experiment (age *m* = 22.95, *sd* = 3.71, range = 18-34). Three participants misinterpreted the instructions and were excluded. The data were collected in the behavioural labs of the Donders Institute, Nijmegen, the Netherlands. The study conformed to the Declaration of Helsinki and was approved by the local ethics committee (CMO2014/288). Written informed consent was obtained from all participants before study enrolment. Colour blindness was an exclusion criterion for this study.

Experiment procedure

In the *no cost* condition, there was no monetary cost of sampling. In the *cost* condition, 5 eurocents were subtracted from the sampling budget for each opened square. In the *not informed* condition, the participant was told that the trustee would not be informed of the information sampling and instead, the trustee believed that the decision to invest was not based on any prior information. In the *informed* condition, the participant was told that the trustee would be informed of the information sampling quantity prior to their decision to reciprocate or defect. We

did not instruct participants about any potential effects that the informed and not informed conditions might have. Each participant completed 60 trials in each of the four conditions, with the four sets of trials being intermixed in random order. The task therefore consisted of 240 trials in total, resulting in a maximum of 6000 sampling decisions, and took approximately 40 minutes to complete. Through parameter recovery, we verified that the number of decisions was large enough to accurately estimate parameters. At the end of the experiment, 3 trials were randomly selected, with the average outcome on those 3 trials computed and paid to the participant (see SI for trial procedure). The outcome of a trust decision was probabilistically determined based on the reciprocation probability r and resulted in a gain of 12 euro if trust was reciprocated, and 0 euro if the trustee defected. The outcome was 6 euro (the initial endowment) if the participant decided not to trust. On *no cost* trials, the entire sampling budget (125 euro cents) was added to the trial outcome, while on *cost* trials, the number of opened boxes times 5 euro cents was subtracted from the sampling budget and the remainder then added to the trial outcome.

Study 2

Experiment procedure

The experimental procedure for study 2 was identical to study 1, with the exception of the reciprocation probabilities. In two (between-subject) conditions, we varied the distribution of reciprocation probabilities. In the positive bias condition, the reciprocation probabilities were: 0.2, 0.4, 0.6, 0.8, and 1. In the negative bias condition the reciprocation probabilities were: 0, 0.2, 0.4, 0.6, and 0.8. Each subject completed a total of 200 trials (40 trials per generative r), resulting in a maximum of 1000 sampling decisions per probability.

Participants

There were 80 participants in total in study 2, 40 participants per condition. In the positively biased condition, the data from one participant was excluded due to a technical malfunction resulting in 39 subjects, 14 males, age: m = 23.15, sd = 2.75, range = 18-34 years). In the negatively biased condition, data from 3 participants were excluded because they did not understand the instructions, resulting in 36 subjects, 11 males, age: m =21.16, sd = 3.13. A one-way ANOVA showed no significant age difference between any of the groups in study 1 and study 2 (p = .091).

Computational Models

General structure of models

A state is determined by the number of open green boxes, n_+, and the number of open red boxes n_-. We consider three models, all with their own variants: An Optimal model, an Uncertainty model, and a Drift Diffusion model (DDM). In each model, the agent computes a decision variable based on n_+ and n_-, and the probability to stop sampling is a noisy function of this decision variable. After the agent stops sampling, the same decision variable determines the probability that the agent trusts. The Optimal and Uncertainty models are based on an evolving posterior distribution over the trustee's reciprocation probability r.

Optimal model

The Optimal model is based on the agent computing expected utility through forward reasoning. It consists of four components: prior beliefs over the r (trustworthiness), an evolving posterior distribution over r, iterative maximization of future expected utility under this posterior distribution, and decision noise temperature. We assume that the prior over r is a beta distribution with parameters α_0 and β_0. The posterior over r is:

$$p(r|n_+, n_-) = \text{Beta}(r; \alpha, \beta), \text{where} \quad \alpha = n_+ \text{ and } \beta = n_- \tag{1}$$

$p(r|\alpha, \beta)$, is then a beta distribution with parameters $\alpha = n_+ + \alpha_0$ and $\beta = n_- + \beta_0$.

A trust outcome can have two values: reciprocation (o = 1, then the investment amount is multiplied by m = 2) or betrayal (o = 0, then the investment amount is multiplied by m = 0). The reciprocation outcome follows a Bernoulli distribution with parameter r:

$$p(\text{outcome} = 1|r) = r \qquad p(\text{outcome} = 0\,|\,r) = 1 - r \tag{2}$$

When trying to predict outcome, the agent does not know r and therefore has to marginalize over r, using the current posterior over r, which is $p(r|\alpha, \beta)$. This gives the conditional distribution of outcome given α and β:

$$p(\text{outcome} = 1|\alpha,\beta) = \int p(\text{outcome} = 1|r)p(r|\alpha,\beta)dr = \int rp(r|\alpha,\beta)dr = \frac{\alpha}{\alpha+\beta} \quad (3)$$

and:

$p(\text{outcome} = 0|\ \alpha,\beta) = \frac{\beta}{\alpha+\beta}$. This is a distribution with mean $\frac{\alpha}{\alpha+\beta}$ and variance $\frac{\alpha\beta}{(\alpha+\beta)^2}$.

We are now ready to define the expected utility of not trusting and trusting. The expected utility of not trusting is $U_0 = 1$, the normalized endowment, which is independent of α and β. The expected utility of trusting should contain the expected amount earned from an investment. However, people are also known to differ in their risk attitude. Specifically in trust games, people may be betrayal-averse [38]. To model such a risk attitude, we subtract a multiplier times the variance of the amount earned from the expected utility. The overall expected utility of trusting, which we denote by U_1, then becomes:

$$U_1(\alpha,\beta) = m \cdot \sum(o|\alpha,\beta) - \lambda m^2 \cdot Var(o|\alpha,\beta) \quad (4)$$

$$= \frac{m\alpha}{\alpha+\beta} - \frac{\lambda m^2 \alpha\beta}{(\alpha+\beta)^2}$$

where λ parametrizes risk attitude, with $\lambda>0$ representing risk aversion and $\lambda<0$ risk-seeking.

At any time, the participant has two possible actions: stopping (a = 0) and sampling (a = 1), except when $t = T+1$, when only $a = 0$ is available. The expected value of a state-action pair, $Q(\alpha, \beta$; action), is given by the Bellman equations[39]. Specifically, if at time t, the participant stops (a = 0), then the expected value of the state (α, β) is the higher of the expected utilities of not trusting and trusting:

$$Q_t(\alpha,\beta; a = 0) = max\{U_0, U_1(\alpha,\beta)\} \quad (5)$$

When $t = T+1$, the value of the state (α, β) is $V_{T+1}(\alpha,\beta) = Q_{T+1}(\alpha,\beta;0)$. At earlier times, V_t is the larger of the two expected utilities:

$$V_t(\alpha,\beta) = max\{Q_t(\alpha,\beta; a = 0), Q_t(\alpha,\beta; a = 1)\} \quad (6)$$

The expected value of a sampling action at time *t* in the state *α, β* by $Q_t(\alpha,\beta; a = 1)$ is:

$$Q_t(\alpha,\beta;\ a = 1) = \frac{\alpha V_{t+1}(\alpha+1,\beta) + \beta V_{t+1}(\alpha,\beta+1)}{\alpha\ +\ \beta} - c \quad (7)$$

where *c* is the subjective cost of a sample.

By starting with the final state (when *n* = *T*) we can apply the equation for the expected value of stopping (equation 3 for *t* = *T*+1, $Q_t(\alpha,\beta; a = 0)$). We can then work our way back in time, to obtain the optimal solution for every possible state (dynamic programming[13]). In the Optimal model, the decision variable, denoted by DV, is the difference between the utilities of sampling and stopping:

$$DV(\alpha,\beta) = Q_t(\alpha,\beta; a = 1) - Q_t(\alpha,\beta; a = 0) \quad (8)$$

The optimal policy would be to sample when the DV is positive; however, we will also introduce decision noise temperature.

Decision noise

The model allows for decision noise through a logistic function:

$$p(sample|\alpha,\beta) = \frac{1}{1 + e^{-\frac{DV(\alpha,\beta)-k}{\tau}}} \quad (9)$$

where DV is the decision variable in the model, *k* is a criterion parameter, and τ is the decision noise (higher *τ* means more noise). Note that noise is not part of the forward reasoning calculations. Instead, it is part of the decision to sample or stop once the optimal solution is derived.

Trust decisions with decision noise

Once the decision to stop sampling has been made, the agent choses to invest using the utility of investing with decision noise:

$$p(invest|\alpha,\beta) = \frac{1}{1+e^{-\frac{U_1(\alpha,\beta)-k}{\tau}}} \tag{10}$$

Uncertainty model

The Uncertainty model is based on the concept of sampling information to reduce uncertainty until a subjective criterion is met. Uncertainty in this model is operationalized as the posterior standard deviation. This model consists of three components: prior belief over the reciprocation probability *r*, an evolving posterior distribution over *r*, and decision noise temperature. As in the Optimal model, we assume that the prior over *r* is a beta distribution with parameters α_0 and β_0. The state at a given time is determined by the number of open boxes (*n*), the number of green boxes (n_+), and the number of red boxes (n_-). The posterior over *r* is then a beta distribution with parameters $\alpha = n_+ + \alpha_0$ and $\beta = n_- + \beta_0$:

$$p(r|\alpha,\beta) = Beta(r;\alpha,\beta) \tag{11}$$

The posterior standard deviation is the decision variable in this model is the standard deviation of this posterior distribution, which is:

$$DV_{Uncertainty}(\alpha,\beta) = std(r|\alpha,\beta) \ = \sqrt{\frac{\alpha\beta}{(\alpha,\beta)^2(\alpha+\beta+1)}} \tag{12}$$

The model allows for decision noise through a logistic function (equation 9).

Drift Diffusion model

In the Drift Diffusion model, the agent keeps track of the absolute difference between positive and negative information and stops sampling when this difference reaches a bound. However, to be consistent with our other models and with most of the value-based decision literature, we will use soft rather than hard bounds.

The decision variable of the "soft" DDM is the absolute difference between these two variables:

$$DV_{\mathrm{DDM}}(n_+, n_-) = |n_+ - n_-| \tag{13}$$

The probability that the agent stops sampling is a logistic function of the difference between this decision variable and a bound b:

$$p(\mathrm{sample}|n_+, n_-) = \frac{1}{1 + e^{-\frac{DV(n_+, n_-) - b}{\tau}}} \tag{14}$$

One can think of this model as a DDM with decision noise. The parameter τ is the temperature of the decision noise. If τ approaches 0, the mapping from DV_{DDM} to the decision becomes deterministic: stop sampling when $|n_+ - n_-| > b$. The presence of decision noise means that the agent sometimes continues sampling even when $|n_+ - n_-| > b$, and sometimes already stops sampling when this inequality is not met.

We consider three versions of the DDM, which differ in the assumptions about the bound b. In the $\mathrm{DDM_s}$ model (where "s" stands for "symmetric"), b is a fixed constant. In the $\mathrm{DDM_a}$ model (where "a" stands for "asymmetric"), b takes three possible values, depending on the sign of $n_+ - n_-$:

$$b = \begin{cases} b_+ & \text{if } n_+ > n_- \\ \dfrac{b_+ + b_-}{2} & \text{if } n_+ = n_- \\ b_- & \text{if } n_+ < n_- \end{cases} \tag{15}$$

In the absence of decision noise, this would be a DDM with asymmetric bounds or with an initial offset[40]. One consequence of this model is that when the difference between positive and negative evidence is close to zero, this model will take a very long time to decide between

trusting or not trusting. To prevent that the agent samples for too long, it may be optimal that the bounds of trusting and not trusting "collapse" towards zero over time[41]. We implemented the collapsing bounds, in the DDMc model (where "c" stands for "collapsing"), by exponentially decaying the bounds towards zero with speed *s* over time, where time is defined as the number of samples *n*:

$$b_c = b * e^{-s*n} \tag{16}$$

where *s* is a free parameter. If *s* is zero, this term is equivalent to a non-collapsing bound.

Trust decisions

Once the decision to stop sampling has been made, the agent choses to trust using the utility of trusting with decision noise:

$$p(\text{trust}|\alpha,\beta) = \frac{1}{1 + e^{-\frac{n_+ - n_- - k_{\text{trust}}}{\tau_{\text{trust}}}}} \tag{17}$$

Note that we allow for a different temperature in the probability of sampling.

Acknowledgments

This study was funded by a European Research Council grant for A.G.S. (#313454). We thank Jeroen van Baar for helpful comments on the manuscript.

References

1 Berg, J., Dickhaut, J. & McCabe, K. Trust, reciprocity, and social history. *Games and economic behavior* **10**, 122-142 (1995).

2 Axelrod, R. & Dion, D. The further evolution of cooperation. *Science* **242**, 1385-1390 (1988).

3 Vives, M.-L. & FeldmanHall, O. Tolerance to ambiguous uncertainty predicts prosocial behavior. *Nature communications* **9**, 2156 (2018).

4 Denrell, J. & March, J. G. Adaptation as information restriction: The hot stove effect. *Organization Science* **12**, 523-538 (2001).

5 Clark, L., Robbins, T. W., Ersche, K. D. & Sahakian, B. J. Reflection impulsivity in current and former substance users. *Biological psychiatry* **60**, 515-522 (2006).

6 Jordan, J. J., Hoffman, M., Nowak, M. A. & Rand, D. G. Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences* **113**, 8658-8663 (2016).

7 Capraro, V. & Kuilder, J. To know or not to know? Looking at payoffs signals selfish behavior, but it does not actually mean so. *Journal of Behavioral and Experimental Economics* **65**, 79-84 (2016).
8 Camerer, C. F. *Behavioral game theory: Experiments in strategic interaction*. (Princeton University Press, 2011).
9 Sanfey, A. G. Social decision-making: insights from game theory and neuroscience. *Science* **318**, 598-602 (2007).
10 Hoffman, M., Yoeli, E. & Nowak, M. A. Cooperate without looking: Why we care what people think and not just what they do. *Proceedings of the National Academy of Sciences* **112**, 1727-1732 (2015).
11 Pérez-Escudero, A., Friedman, J. & Gore, J. Preferential interactions promote blind cooperation and informed defection. *Proceedings of the National Academy of Sciences* **113**, 13995-14000 (2016).
12 Hilbe, C., Hoffman, M. & Nowak, M. A. Cooperate without looking in a non-repeated game. *Games* **6**, 458-472 (2015).
13 Bellman, R. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences* **38**, 716-719 (1952).
14 Simon, H. A. Theories of bounded rationality. *Decision and organization* **1**, 161-176 (1972).
15 Gigerenzer, G. & Goldstein, D. G. Reasoning the fast and frugal way: models of bounded rationality. *Psychological review* **103**, 650 (1996).
16 Ratcliff, R. & Rouder, J. N. Modeling response times for two-choice decisions. *Psychological Science* **9**, 347-356 (1998).
17 Chang, L. J., Doll, B. B., van't Wout, M., Frank, M. J. & Sanfey, A. G. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive psychology* **61**, 87-105 (2010).
18 Aimone, J. A., Houser, D. & Weber, B. in *Proc. R. Soc. B.* 20132127 (The Royal Society).
19 Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J. & Friston, K. J. Bayesian model selection for group studies. *Neuroimage* **46**, 1004-1017 (2009).
20 Hill, C. A. *et al.* A causal account of the brain network computations underlying strategic social behavior. *Nature neuroscience* **20**, 1142 (2017).
21 Wilson, R. K. & Eckel, C. C. Judging a book by its cover: Beauty and expectations in the trust game. *Political Research Quarterly* **59**, 189-202 (2006).
22 Delgado, M. R., Frank, R. H. & Phelps, E. A. Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature neuroscience* **8**, 1611 (2005).
23 Fareri, D. S., Chang, L. J. & Delgado, M. R. Effects of direct social experience on trust decisions and neural reward circuitry. *Frontiers in neuroscience* **6**, 148 (2012).
24 Baumeister, R. F., Bratslavsky, E., Finkenauer, C. & Vohs, K. D. Bad is stronger than good. *Review of general psychology* **5**, 323 (2001).
25 Skowronski, J. J. & Carlston, D. E. Negativity and extremity biases in impression formation: A review of explanations. *Psychological bulletin* **105**, 131 (1989).
26 Fiske, S. T. Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of personality and Social Psychology* **38**, 889 (1980).
27 Mende-Siedlecki, P., Baron, S. G. & Todorov, A. Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. *Journal of Neuroscience* **33**, 19406-19415 (2013).

28 Chase, V. M., Hertwig, R. & Gigerenzer, G. Visions of rationality. *Trends in cognitive sciences* **2**, 206-214 (1998).
29 Nelson, J. D. Naïve optimality: Subjects' heuristics can be better motivated than experimenters' optimal models. *Behavioral and Brain Sciences* **32**, 94-95 (2009).
30 Griffiths, T. L., Lieder, F. & Goodman, N. D. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science* **7**, 217-229 (2015).
31 Gershman, S. J., Horvitz, E. J. & Tenenbaum, J. B. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* **349**, 273-278 (2015).
32 Bradley, B. P., Mogg, K., Millar, N. & White, J. Selective processing of negative information: Effects of clinical anxiety, concurrent depression, and awareness. *Journal of abnormal psychology* **104**, 532 (1995).
33 Hauser, T. U., Moutoussis, M., Dayan, P. & Dolan, R. J. Increased decision thresholds trigger extended information gathering across the compulsivity spectrum. *Translational psychiatry* **7**, 1296 (2017).
34 Hauser, T. U. *et al.* Increased decision thresholds enhance information gathering performance in juvenile Obsessive-Compulsive Disorder (OCD). *PLoS computational biology* **13**, e1005440 (2017).
35 King-Casas, B. & Chiu, P. H. Understanding interpersonal function in psychiatric illness through multiplayer economic games. *Biological psychiatry* **72**, 119-125 (2012).
36 Lazarus, S. A., Cheavens, J. S., Festa, F. & Rosenthal, M. Z. Interpersonal functioning in borderline personality disorder: a systematic review of behavioral and laboratory-based assessments. *Clinical Psychology Review* **34**, 193-205 (2014).
37 Hula, A., Vilares, I., Lohrenz, T., Dayan, P. & Montague, P. R. A model of risk and mental state shifts during social interaction. *PLoS computational biology* **14**, e1005935 (2018).
38 Bohnet, I. & Zeckhauser, R. Trust, risk and betrayal. *Journal of Economic Behavior & Organization* **55**, 467-484 (2004).
39 Bellman, R. A Markovian decision process. *Journal of Mathematics and Mechanics*, 679-684 (1957).
40 Mulder, M. J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W. & Forstmann, B. U. Bias in the brain: a diffusion model analysis of prior probability and potential payoff. *Journal of Neuroscience* **32**, 2335-2343 (2012).
41 Tajima, S., Drugowitsch, J. & Pouget, A. Optimal policy for value-based decision-making. *Nature communications* **7**, 12400 (2016).