Cognition 231 (2023) 105317

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original articles

Rational information search in welfare-tradeoff cognition

Tadeg Quillien

School of Informatics, University of Edinburgh, United Kingdom

ARTICLE INFO

Dataset link: https://osf.io/jtavm

Keywords:
Social cognition
Active learning
Computational modeling
Evolutionary psychology
Theory of Mind

ABSTRACT

One of the most important dimensions along which we evaluate others is their propensity to value our welfare: we like people who are disposed to incur costs for our benefit and who refrain from imposing costs on us to benefit themselves. The evolutionary importance of social valuation in our species suggests that humans have cognitive mechanisms that are able to efficiently extract information about how much another person values them. Here I test the hypothesis that people are spontaneously interested in the kinds of events that have the most potential to reveal such information. In two studies, I presented participants (Ns = 216; 300) with pairs of dilemmas that another individual faced in an economic game; for each pair, I asked them to choose the dilemma for which they would most like to see the decision that the individual had made. On average, people spontaneously selected the choices that had the potential to reveal the most information about the individual's valuation of the participant, as quantified by a Bayesian ideal search model. This finding suggests that human cooperation is supported by sophisticated cognitive mechanisms for information-gathering.

1. Introduction

Your friend gave you an expensive ticket for a concert of their favorite artist as a birthday gift. You know that the date of the concert happens to coincide with a conference that she was unexpectedly asked to attend. You might be curious about whether she had planned to go herself but could not make it and recycled the item by gifting it to you.

Why would one be curious about this? The origin of the ticket does not change its intrinsic value, because the artist's performance will be the same in any case. Intuitively, one reason to care is that the ticket's provenance contains some information about how much your friend cares about you. If she bought this expensive item for the specific purpose of gifting it to you, you learn that she is willing to incur high costs for your benefit. You learn no such thing if the gift was simply recycled.

Here I examine the hypothesis that human curiosity is well-designed to gather that sort of information. That is, we are good at looking for information about how much someone values our welfare.

1.1. The psychology of social valuation

Successful interaction with other people requires that we represent their mental states, such as their preferences and their desires, in order to predict and influence their behavior. In particular, one dimension along which it is crucial to evaluate other people is the extent to which they value us.

People often have to make tradeoffs between their own *welfare* and that of others. That is, some of the actions we contemplate may benefit us, but generate a cost to someone else. Or, some actions may benefit others but be costly for us. For example, if Alice is pondering whether she should offer Bob a ride to the airport on a day he needs it, she will do so to the extent that she thinks the benefit to Bob offsets the cost to her. As such, the human mind might contain internal variables that regulate the kinds of trade-offs we make between a person's welfare and ours. That is, when Alice decides whether to help Bob, she might implicitly consult a set of person-specific variables that regulate how much she should help Bob (Delton, 2010; Delton & Robertson, 2016; Sell, 2005; Tooby, Cosmides, Sell, Lieberman, & Sznycer, 2008). As a simple example, if Alice is considering whether paying a cost c to deliver a benefit b to Bob, she might consult an internal variable specifying the ratio $\frac{b}{c}$ above which she should help. Here I call these variables "welfare trade-off parameters" (WTPs).[1] In folk-psychological terms, the people who have favorable welfare trade-off parameters toward us are the people who like us, respect us, are willing to help us, etc.

Other people vary in how much they value our welfare. It is also possible to influence the weight that a given person puts on our welfare,

E-mail address: tadeg.quillien@gmail.com.

[1] The psychology of social valuation is often illustrated with a toy model where Alice has a single WTP towards Bob, her Welfare-Tradeoff Ratio (WTR), as in the example above. The experiments I report here have a simple enough structure that the single-parameter WTR model is an adequate model of human behavior (see Delton, 2010), but a full cognitive model would require more parameters to explain behavior in more complex settings, for example to capture the extent to which people are sensitive to variation in the cost of a helpful action (Fisman, Kariv, & Markovits, 2007).

https://doi.org/10.1016/j.cognition.2022.105317
Received 27 May 2022; Received in revised form 23 August 2022; Accepted 25 October 2022
Available online 23 November 2022

for example by making threats and promises, or being generous to them (Lim, 2012; Quillien, 2020; Sell, Tooby, & Cosmides, 2009). This suggests that it would be highly adaptive to care about how much another person values us. When deciding whether to avoid that person or be friends with them, whether to trust them, how much to invest in a relationship with them, etc, our decision might be guided, in part, by how much we think they value us. In the next section I review evidence that this is the case.

1.2. The role of inferences about social valuation

Evolutionary considerations suggest that inferences about Welfare-tradeoff parameters (WTPs) may be the most fundamental criterion by which we evaluate others (Eisenbruch & Krasnow, 2022; Fiske, Cuddy, & Glick, 2007; Qi & Vul, 2022; Sell, 2005; Tooby & Cosmides, 2008). Humans are an intensely social species: our hunter–gatherer ancestors relied on each other for help in many domains including food production, food sharing, childrearing, and warfare (Gurven, 2004; Hrdy, 2011; Kappeler & Van Schaik, 2006; Tooby & Cosmides, 2010; Wrangham, 1999). For example, adults who were injured or ill depended on others for survival (Gurven, 2004; Sugiyama, 2004; Tooby & Cosmides, 1996). Sugiyama (2004) finds, in a contemporary forager–horticulturalist population, that about 65% of men have experienced a disability lasting for at least a month, and would probably have died without provision of food from community members. Therefore, it was essential for ancestral humans to ensure that other community members valued their welfare highly (Delton & Robertson, 2012; Tooby et al., 2008).

Evaluating the WTPs of another person is especially important because it often determines how we should assess the other characteristics of that person (Fiske et al., 2007). For example, the fact that someone has strong fighting skills should make you like them more if they value you, but it should make you like them less if they are your enemy. Recently, Eisenbruch and Krasnow (2022) have argued that facts about the *statistical distributions* of WTPs can warrant their prioritization in social evaluation.[2]

A large empirical literature suggests that WTPs are indeed prioritized in social evaluation.

A number of studies have looked at how people choose their partners in simple economic games. In such studies, participants typically interact with several different partners. These partners vary in their WTPs (i.e. some share a larger portion of their endowment than others), and in their "productivity" (i.e. some give a larger absolute amount of money to the participant). When participants have to choose which partner they want to continue interacting with in subsequent games, they typically privilege cues of WTPs over cues of raw productivity. For example, a partner who shares 20% of his $0.50 endowment with the participant gives more money ($0.10) to the participant than a partner who shares 50% of his $0.10 endowment ($0.05), yet the latter demonstrates that she values the participant more. Participants tend to prefer to play with the partner that seems to value them more, and often do so even in situations where this is not the economically rational option (Eisenbruch & Roney, 2017; Hackel, Doll, & Amodio, 2015; Hackel, Mende-Siedlecki, & Amodio, 2020; Lim, 2012; Raihani & Barclay, 2016; see also Delton & Robertson, 2012).

Research on social emotions also suggests that people spontaneously make inferences about the WTPs of others (Sznycer, Sell, & Lieberman, 2021). Anger may serve to communicate to the target that they do not value us highly enough (Sell, 2005; Sell et al., 2009). Gratitude may signal to the target that we acknowledge that what they did reveals that they value us highly (Lim, 2012; Smith, Pedersen, Forster, McCullough, & Lieberman, 2017; Sznycer et al., 2021). In support of these theories, researchers have found that the actions of others elicit our anger or gratitude to the extent that they reveal information about the agent's WTPs toward us (Lim, 2012; Quillien, 2021; Sell et al., 2017; see also Monroe, 2020; Tesser, Gatewood, & Driver, 1968; Yu, Gao, Zhou, & Zhou, 2018). For example, Sell et al. (2017) asked participants to read vignettes in which a perpetrator inflicts a cost on the participant in order to get a benefit (for instance, cut in line at a public telephone booth in order to communicate an urgent message). They manipulated the benefit to the perpetrator, the cost to the participant, as well as the perpetrator's intention. Participants were most angry at perpetrators whose actions revealed that they did not value the participant (they intentionally inflicted a large cost on the participant, in order to gain a trivial benefit). This result held in all cultures surveyed, including a small-scale Amazonian society (Sell et al., 2017).

Additionally, a large literature in social psychology finds that (i) people evaluate others primarily along two dimensions, their "warmth" and their "competence", and that (ii) they tend to prioritize warmth (Fiske et al., 2007; Wojciszke, 2005). The warmth dimension captures traits such as friendliness, morality, trustworthiness and helpfulness, while the competence dimension encompasses traits such as creativity, skill, and efficacy. Arguably, the warmth dimension captures traits that index an individual's WTPs, while the competence dimension captures traits that underlie an individual's ability to generate costs and benefits (Eisenbruch & Krasnow, 2022).

Warmth information usually has a larger influence than competence on impression formation (Fiske et al., 2007; see also Eisenbruch & Krasnow, 2022). For instance, when people are asked to rate their acquaintances on ten warmth-relevant and ten competence-relevant traits, and also give their overall impression of the acquaintance, ratings on warmth-relevant traits are a better predictor of overall impression (Wojciszke, Bazinska, & Jaworski, 1998).

Recently it has been proposed that the ability to represent welfare-tradeoff parameters is the cornerstone of early social cognition (Powell, 2022). From an early age, children draw sophisticated inferences about an agent's valuation of others — for example, when they see two agents fail to help another agent, children tend to infer that the less competent agent, for whom the cost of helping would have been higher, is nicer (Jara-Ettinger, Tenenbaum, & Schulz, 2015).

1.3. The current research

Cognitive scientists have remarked that motivational systems co-evolve with representational systems (Delton & Sell, 2014; Tooby, Cosmides, & Barrett, 2005). For example, a fear of predators is useless without the ability to detect predators (Barrett, 2005), and a motivation to help kin members cannot evolve without mechanisms for recognizing kin (Lieberman, Tooby, & Cosmides, 2007; Sznycer, De Smet, Billingsley, & Lieberman, 2016a). Many aspects of human social emotions and social behavior seem to be guided by inferences about how much a given person values our welfare. This suggests that the human mind is equipped with cognitive mechanisms that can efficiently construct accurate representations of the welfare-tradeoff parameters of others.

[2] First, WTPs exhibit higher between-agent variance than competence does. Some people (e.g. your rivals) actively hate you, while others (e.g. your mother) value you highly; by contrast competence tends to be more evenly distributed across people, especially in ancestral environments. Second, WTPs exhibit lower within-agent variance than competence does (i.e., WTPs exhibit higher stability across domains). If someone is sharing food with you, it is likely the she would also offer you shelter when you need it. By contrast, a good hunter is not necessarily a good carpenter. In conjunction, these two statistical facts make WTP information a prime target for information acquisition, and an important factor for partner choice. Because of their high between-agent variance, the WTPs of a new person are the feature you are initially the most uncertain about, so it is the one for which new information is most valuable. Also, because of their low within-agent variance, information about WTPs that you glean from a single action (e.g. someone sharing food with you) is likely to be highly diagnostic of the person's future behavior, so it should be weighed highly when choosing a partner (Eisenbruch & Krasnow, 2022).

This hypothesis is bolstered by recent advances in our understanding of commonsense psychology. There is increasing evidence that people might predict and explain the behavior of other agents by representing them as expected-utility maximizers (Gates, Callaway, Ho, & Griffiths, 2021; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Jern, Lucas, & Kemp, 2017; Lucas et al., 2014). That is, we assume that other agents behave in approximately rational ways given their beliefs and preferences. Modeling other people as expected-utility maximizers is a very flexible way of making useful inferences and predictions, and humans make such inferences in an approximately rational way, even from an early age (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Jara-Ettinger et al., 2016; Liu, Ullman, Tenenbaum, & Spelke, 2017).

This body of work suggests that we might understand how an agent makes welfare-tradeoffs by assuming that their utility function takes into account the costs and benefits that their actions generate for others. Modeling other people in this way would allow us to infer the weight they put on our welfare by observing what they do. For example, suppose we observe someone knowingly inflicting a large cost on us as a means of getting a trivial benefit. The only way to rationalize their behavior is to assume that they do not put a large weight on our welfare, because otherwise they would not consider that the trivial benefit to them offsets the large cost on us. In sum, if we are aware that people's preferences can incorporate the welfare of other agents, then we might be able to infer their WTPs by using the inference algorithms that we use to infer people's preferences in general.

Is the human mind well-equipped to construct representations of the WTPs of others? Here I test one specific aspect of this hypothesis, by looking at people's information-gathering behavior. One expects that people will be curious about events that have the potential to reveal a lot of information about how much someone values them. All else equal, if people have the opportunity to observe either one of two events, they will be more likely to observe the event that potentially contains the most information about someone's welfare tradeoff parameters toward them.

Existing work has found that people are particularly curious about the WTPs of others, relative to other traits; for example when asked which traits they would most like to learn about a person in order to form an overall impression of that person, people are more likely to ask for warmth-relevant traits such as *fair*, *generous*, *righteous*, *sincere*, than competence-relevant traits such as *clever*, *foresighted*, *ingenious*, *intelligent* (Wojciszke et al., 1998). Even within the domain of warmth-related traits, people are more interested in traits relevant to morality than traits related to sociability, such as *friendliness* and *likeability* (Brambilla, Rusconi, Sacchi, & Cherubini, 2011). Simulation studies have also shown that one expects natural selection to design agents who are more interested in gathering information about another agent's WTPs, compared to other traits of that agent (Eisenbruch & Krasnow, 2022).

Here I am interested in a slightly different question: assuming that people are motivated to gather information about WTPs, are they able to compute the amount of information they would gain about someone's WTPs if they observe that person making a given decision?

Intuitively, people are at least somewhat competent at this task. We know that looking at whether someone will hold the door for us will not give us that much information about how much they value us, compared to learning whether they are willing to donate us a kidney. But how sensitive are we to fine-grained variation in the potential information content of an event?

To address this question, I compare people's choices in a simple data selection task to a normative model. To develop this normative model, I draw inspiration from work on the cognitive psychology of information-gathering.

1.4. The psychology of information-gathering

Selecting information is a ubiquitous problem for most creatures. There is an almost infinite number of things we could observe, experiments we could perform, questions we could ask, places we could direct our gaze to…, but some of them are more likely to give us the information we need. How do we choose?

Early research has painted a pessimistic picture of human information-gathering. Even in simple logic puzzles, people are extremely bad at identifying the information that would help them get the right answer (Wason, 1960, 1968). For instance, in the Wason selection task, most people are unable to identify the information that would allow them to falsify a simple logical rule of the form "If P, then Q" (Wason, 1968). Yet, people's failures at such simple tasks need not imply that the human mind is not well-designed for information-gathering. It is unlikely that natural selection designed the human mind for the ability to solve abstract logical problems (Cosmides, 1989), and therefore participants in these experiments may be implicitly solving a different problem. Indeed, if we assume that participants in the Wason selection task are implicitly trying to acquire information about causal regularities in the world, their typical pattern of answer reflects the optimal one (Oaksford & Chater, 1994; see also Hendrickson, Navarro, & Perfors, 2016; Navarro & Perfors, 2011). In addition, participants often give the correct answer to the selection task when its content triggers mechanisms designed to solve ancestrally-relevant problems (Cosmides, 1989; Cosmides & Tooby, 2005; Gigerenzer & Hug, 1992).

In order to assess the rationality of people's data selection choices, researchers often compare them to the choices prescribed by a normative model (Klayman & Ha, 1987; Nelson, 2005; Oaksford & Chater, 1994; Tsividis, Gershman, Tenenbaum, & Schulz, 2014). Normative models of data selection typically use the framework of Optimal Experimental Design (OED), which draws on insights from statistics and information theory. OED models are structured around two main components: an ideal observer model, which specifies the inferences that an agent should make, given an observation, and a measure of how much information such an observation would give the ideal observer (Liefgreen, Pilditch, & Lagnado, 2020; Nelson, 2005; although see Dubey & Griffiths, 2020 for another type of normative model).

Using such techniques, researchers have found that the mind hosts many systems that bear the signature of optimal or near-optimal information-gathering. For example, people make eye movements that are optimized for gathering visual information efficiently (Cain, Vul, Clark, & Mitroff, 2012; Najemnik & Geisler, 2005; Nelson & Cottrell, 2007; Peterson & Eckstein, 2012, 2013). They make interventions that are effective for learning about the causal structure of a system (Bramley, Lagnado, & Speekenbrink, 2015; Jiang & Lucas, 2021; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003), and the physical properties of novel objects (Bramley, Gerstenberg, Tenenbaum, & Gureckis, 2018). They seek information that is helpful for categorization (McKenzie, 2006; Nelson, 2005; Nelson, McKenzie, Cottrell, & Sejnowski, 2010). Even from a young age, they tend to ask questions that effectively narrow down the space of possible hypotheses (Nelson, Divjak, Gudmundsdottir, Martignon, & Meder, 2014; Ruggeri & Lombrozo, 2015; see also Rothe, Lake, & Gureckis, 2018).

Selecting optimal queries is a computationally complex task — therefore it is plausible that the human mind uses resource-rational strategies to approximate the optimal solution to information search problems (Coenen, Nelson, & Gureckis, 2019; Liefgreen, Pilditch & Lagnado, 2020). Indeed, researchers have found evidence that in many settings, people probably use heuristics that approximate the optimal search strategy given the appropriate context (Bramley et al., 2015; Gelpi, Saxena, Lifchits, Buchsbaum, & Lucas, 2021; Markant & Gureckis, 2014; Markant, Settles, & Gureckis, 2016; Wu, Meder, Filimon, & Nelson, 2017).

In sum, empirical research suggests the following picture of human information search. Identifying the most informative query one could

make is computationally complex, and people do not systematically succeed. Indeed, sometimes they perform much worse than chance (Wason, 1968), perhaps because the task (e.g. a logical puzzle) is not natural and people implicitly substitute a more natural problem to it. Nonetheless, in some settings, people can make choices that are close to the normative ideal.

I predict that information-gathering about welfare-tradeoff parameters is one such setting. Given that the representation of WTPs is an essential part of social cognition, a task where one can select WTP-relevant information should not feel like an abstract puzzle to participants. Instead it should elicit mechanisms that are designed to reliably extract this information.

1.5. Overview of the task and normative model

In the current experiments, I use a data selection task, set in the context of a simple economic game that participants play with a partner. In this game, the partner has the opportunity of behaving either generously or selfishly. I show participants pairs of trials of the game that their partner has played. Participants can see the choice that was available to their partner in each trial, but not what their partner actually did. I ask them for which trial they would most like to see what their partner did.

Note that the game does not involve third-parties, so the current experiment does not strictly distinguish between the task of inferring how much your partner values your welfare in particular, and the task of inferring how much your partner is willing to trade her welfare against others' in general. I return to this point in the General Discussion.

The allocation task that participants play with their partner is the Welfare-Tradeoff Task (WTT; Delton, 2010). In a trial of the WTT, the *dictator* has to choose between two possible allocations of money between herself and the *recipient*:

-$X for the dictator, $0 for the recipient

or

-$0 for the dictator, $Y for the recipient.

Depending on the values of X and Y, watching your partner play the game as a dictator can be more or less informative about how much she values your welfare.

For example, if your partner has a choice between $20 for herself and $1 for you (i.e. X=20, Y=1), you can predict with high confidence that she will take the $20 for herself. If you observe her doing so, you will not have learnt much about her. This is because, even if she valued you highly, she probably would still have taken the $20 for herself (she probably does not think that a dollar to you is worth more than twenty times a dollar to herself). Therefore, observing what your partner did in this round of the game does not contain a lot of potential information.

By contrast, imagine that your partner has a choice between $20 for herself and $30 to you. It is less easy to predict what she will do. Both possible outcomes (that she takes the $20 for herself, or that she lets you have the $30) are plausible. Therefore, either way you will have learnt something. If you see your partner make the self-serving allocation, you will decrease your estimate of the weight she puts on your welfare. If you see her make the generous allocation, you will increase your estimate.

Therefore, intuitively, if you were able to see what your partner did in only one of these situations ($20 for her vs $1 for you, or $20 for her vs $30 for you), you should be more interested in inquiring about the second situation. Below I explain how one can formalize this intuition using standard mathematical tools, to systematically quantify how interesting it is to observe a given decision. For now I give a qualitative overview of these techniques, accessible to a non-mathematically-inclined reader, and I give a more formal treatment in the 'Computational modeling' section.

1.5.1. Inference

The amount of information you gain from an observation depends on the inference you make on its basis. Therefore, in order to define the value of information we first need a model of how people make inferences about how much a partner values them, when they see the partner make a decision.

I assume that people make these inferences by using *Bayes' rule* to invert a *generative model* of how people make decisions in the WTT.

First, people have a generative model of how people behave in the WTT. This means that people have some expectations about how other agents tend to play the game. I present the formal version of this generative model in the methods; for now it is enough to say that it formalizes three relatively mild assumptions:

-agents are more likely to make the generous allocation as the opportunity cost of doing so decreases. That is, if you have a choice between $X for you and $Y for the other person, you are *increasingly* likely to be generous as X *decreases*.

-different agents can assign different weights to the welfare of their partner, and a given agent can assign a different weight to the welfare of different partners. That is, some agents have a greater tendency for generosity than others, and a given agent may value some people more than others.

-agents are not always perfectly consistent across their decisions. It is possible to give the exact same dilemma to an agent at different times, and see the agent make a different decision the second time than the first. This is especially likely to happen for a dilemma where the agent is nearly indifferent between the options at her disposal.

Second, when people observe an agent make a decision (for instance, decide to take $20 for herself instead of letting her partner have $30), they use *Bayes' rule* to update their estimate of the weight that the agent puts on their welfare. Bayes' rule is a simple theorem in probability theory that implies a normatively correct way of updating one's beliefs when observing new relevant data.

1.5.2. Information value

It is possible to compute how much information you gain from an observation, assuming that you make inferences in the way specified in the previous section.

If you see your partner make a decision (e.g. she takes $20 instead of letting you have $30), how much information do you gain about how much she values you? This can be quantified as the Kullback–Leibler divergence (Kullback & Leibler, 1951) between your belief after the observation and your belief before the observation.[3]

Conceptually, the KL divergence measures how much your belief about an aspect of the world changes in response to an observation. Fig. 1 explains the intuition for what KL divergence is measuring in the current context. Here the space of all possible hypotheses is continuous. There are an infinity of possible hypotheses about Alice's Welfare-Tradeoff Ratio (WTR) towards Bob (it could be .4012, .4013,

[3] Several other measures of the information value of an observation exist (see Nelson, 2005). Two early measures, *Bayesian diagnosticity* and *log-diagnosticity* are now seen as unsatisfactory normative models (Nelson, 2005). *Probability gain* measures how much an observation increases one's likelihood of making a correct categorization decision, and is therefore not well suited for the current setting in which the ideal search model infers the value of a continuous variable; in addition, probability gain sometimes assigns negative value to observations, which is undesirable for a normative model (Liefgreen, Pilditch & Lagnado, 2020). *Information gain* quantifies the value of an observation differently than KL, but both measures give identical values for the expected information value of a query (Oaksford & Chater, 1996), so they make identical predictions here. *Impact* is conceptually similar to KL, in that it measures how much an observation changes one's belief; empirically, in the current setting the ideal search model made identical predictions whether it computed information value using KL or Impact.

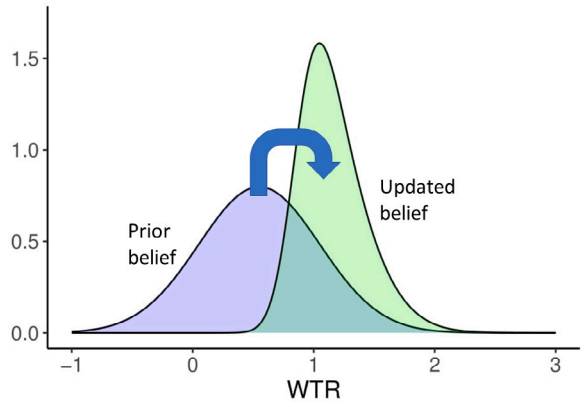

Fig. 1. Conceptually, the KL divergence quantifies how much one must move the probability distribution corresponding to the observer's prior belief to obtain the belief that has been updated by the observation. The greater the divergence of the updated belief from the prior belief, the higher the information content of the decision. Y-axis: probability density.

.4014, etc.), and this continuum of hypotheses is plotted on the *x*-axis. The blue curve is the observer's prior belief about Alice's WTR; it is a probability distribution over all possible values for Alice's WTR. The green curve is the observer's posterior belief after having seen a decision by Alice (here, a decision where she chose the generous option). Intuitively, the KL divergence measures how much you have to 'move' the blue curve in order to make it into the green curve.

1.5.3. Expected information value

In the current task, people have to choose which decision to observe. That is, they have to estimate how much information they will get from observing the outcome of a given decision, *before knowing how their partner actually behaved*. In other words, people need to compute the value of a *query*, rather than the value of an *observation*. The value of a query is simply the expected information value across the possible observations that could result from that query. Maybe your partner will be selfish, maybe your partner will be generous. To compute the value of the query, you have to compute a weighted average of the information value of these two possible observations. The weight of each observation is simply your estimate of the probability of this observation.

1.5.4. Choice

After computing the expected information value of two different queries, people need to decide which query to make. A perfectly rational agent would simply always select the query for which she computed the highest expected information value. However, because of various sources of noise, participants probably will not always select the best option. Instead, one can predict that people will have a statistical tendency to select the queries with the highest expected information value. Crucially, the higher the difference in expected information value between the two queries, the most likely people will be to select the 'best' one.

1.5.5. Summary

Taken together, the steps described above constitute an "ideal search model". An agent making queries according to such a model would conform to normative principles for how to acquire data that are maximally informative.

In the next section, I describe the ideal search model more formally. Then I report two studies designed to test the hypothesis that human behavior is well-approximated by the ideal search model.

2. Computational modeling

Alice and Bob are playing an economic game, the Welfare Tradeoff Task (WTT), with Alice as dictator and Bob as recipient. Given a trial of the WTT, can we, before seeing what Alice did, estimate how much information we would learn about how much she values Bob's welfare, if we observed her decision in that trial? The ideal search model is designed to solve this task.

The WTT is a two-player game with a dictator and a recipient. In a trial of the WTT, if Alice is the dictator and Bob is the recipient, Alice must choose between the two alternatives:

Alice receives $\$\pi_{alice}$ and Bob receives nothing

Or

Bob receives $\$\pi_{bob}$ and Alice receives nothing

The dictator plays several trials of the game. Across trials, the value of $\$\pi_{alice}$ varies, while $\$\pi_{bob}$ remains almost constant. The dictator is told that only one trial will be randomly selected to be paid out, and that therefore she should treat each trial as if it was the only one.

The first component of the ideal search model is an ideal observer model. The ideal observer model updates its belief about the weight that Alice puts on Bob's welfare, after observing Alice's decisions. Elsewhere (Quillien, Tooby, & Cosmides, in preparation), I report data suggesting that this model is a good approximation for how people make inferences about someone's WTR after observing that person's choices.

2.1. Ideal observer model

The ideal observer model relies on a causal model of how Alice makes decisions. This causal model holds that the weight that Alice puts on Bob's welfare can be represented by a single welfare-tradeoff parameter, which we call her Welfare-Tradeoff Ratio (WTR) towards Bob. While this assumption is unrealistic, the WTT is simple enough that a single-parameter model is a good model of how people actually play the task. Specifically, empirical data (Delton, 2010) suggests that when playing the WTT, Alice tries to maximize her expected utility, given by Eq. (1):

$$U_{alice} = \pi_{alice} + \text{WTR}_{alice->bob} * \pi_{bob} \quad (1)$$

Intuitively, Alice's WTR toward Bob is what Alice considers to be the 'exchange rate' between her welfare and Bob's. For example, if WTR = 1/2, she values Bob's welfare half as much as her own. The decision rule that follows from this utility function is that Alice allocates the money to Bob if:

$$\text{WTR}_{alice->bob} > \frac{\pi_{alice}}{\pi_{bob}}$$

I also assume that Alice observes a noisy value of the payoffs in each trial. Specifically, for each trial with $\phi = \frac{\pi_{alice}}{\pi_{bob}}$ she observes a noisy value $\hat{\phi} = \phi + \epsilon$, where the noise ϵ is drawn from a normal distribution with mean 0 and variance σ^2_ϕ. This constraint makes her choices non-deterministic, and models the fact that humans are not always perfectly consistent in their behavior when they make welfare-tradeoffs (Delton, 2010; Fisman et al., 2007).[4]

Using this causal model, one can compute the probability that Alice makes a given decision ('Give' or 'Take') in a specific trial of the WTT, given her WTR toward Bob. Specifically, we have:

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

$$P(\text{Take}|\text{WTR}, \phi) = 1 - P(\text{Give}|\text{WTR}, \phi)$$

[4] An alternative way to introduce stochasticity in Alice's choices would be to assume that her choices are the output of a softmax function, where a 'temperature parameter' regulating the stochasticity of the choices would play the same role as the variance parameter used here.

where ϵ is the observation noise with which Alice observes the value of ϕ.

The ideal observer's belief in Alice's WTR is not a point estimate, but a probability distribution. We write this probability distribution as $P(\text{WTR})$: it is a function that assigns a relative probability density to each possible WTR that Alice could have toward Bob. Given this belief, the ideal observer can compute the probability that Alice will Give or Take in a given trial of the WTT. It does that according to the law of total probability, by computing a weighted sum of the likelihood term $P(\text{decision}|\text{WTR}, \phi)$ for different WTRs, where each possible WTR is weighted according to its probability. Formally, we write this as:

$$P(\text{decision}|\phi) = \int P(\text{decision}|\text{WTR}, \phi)P(\text{WTR})\, d\text{WTR}$$

When observing Alice make a decision in a trial of the WTT with payoff ratio ϕ, the ideal observer updates his belief in Alice's WTR via Bayes' rule:

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

where $P(\text{WTR})$ denotes the model's prior belief in Alice's WTR.

Algorithmically, I used grid approximation to implement the ideal observer. The R code for the implementation is available at the Open Science Framework.[5]

2.2. Expected information value

The information value of an observation d is the Kullback–Leibler (KL) divergence between the ideal observer's posterior belief about Alice's WTR and its prior belief:

$$U(d) = \text{KL}(P(\text{WTR}|d) \parallel P(\text{WTR}))$$

$$= \int_{-\infty}^{\infty} P(\text{WTR}|d) \log\left(\frac{P(\text{WTR}|d)}{P(\text{WTR})}\right) d\text{WTR}$$

The expected information value of a query is simply the weighted mean of the information value of its possible outcomes:

$$\text{EIV} = \sum_i Pr(d_i)U(d_i)$$

$$= \text{KL}(\text{Take})Pr(\text{Take}) + \text{KL}(\text{Give})Pr(\text{Give})$$

2.3. Choice

The ideal search model must choose, among two trials A and B, which one to observe. The optimal strategy is to always pick the trial with the highest expected information value. However, this might not be the best choice for comparison with human data. Even assuming that humans can compute the expected information value of a trial, one does not expect them to always select the trial with the highest information value (because of inattention, noise in neural processing, exploratory behavior, etc.). Instead one expects them to select a trial with a probability that is a function of its relative expected information value.

To model this, when choosing between trials A and B the ideal search model selects trial A with probability:

$$Pr(A) = \frac{e^{\beta I(A)}}{e^{\beta I(A)} + e^{\beta I(B)}}$$

where $I(X)$ is the expected information value of X, and β is an 'inverse temperature' parameter, determining the amount of stochasticity in the selection (for $\beta = 0$, the model selects randomly; the higher the value of β the closer the model is to always selecting the most valued option), whose value will be fit to the human data.

2.4. Alternative models

In addition to the ideal search model, I tested three alternative computational models of data selection. All models were built on top of the ideal observer, but used its predictions in different ways.

The first model, 'optimal search without updating' was a 'lesioned' version of the ideal search model, which works in the same way, with the exception that it does not update its belief about Alice's WTR after observing her decisions. This model can be seen as formalizing the null hypothesis that the manipulation of participants' prior beliefs has no effect.

The 'confirmation' model is inspired by research showing that people sometimes employ a 'positive test' strategy (Coenen, Rehder, & Gureckis, 2015; Klayman & Ha, 1987; Wason, 1968), asking to observe events that have the potential to confirm their current hypothesis. In the current context, positive testing does not systematically lead to the maximization of expected information gain, so I included this model to test that people track expected information gain more closely than under a positive testing strategy. I also tested the opposite approach, a 'Falsification' model, which asks to observe events that have the potential to disconfirm its current hypothesis. I tested the Falsification model because is a natural complement to the Confirmation model. What counts as 'Falsification' or 'Confirmation' in the context of the estimation of a continuous value is somewhat ambiguous, so there could in principle be many ways that one could operationalize such a strategy. Here I choose to test the following very simple implementation.

In the 'confirmation' model, the agent selects the trial where the partner is most likely to act the same as she did before. If the partner's first decision was 'Take', the model requests to observe the trial in which it predicts that the partner is most likely to Take; if the partner's first decision was 'Give', the model requests to observe the trial in which it predicts that the partner is most likely to Give.

The 'Falsification' model does the opposite: it tries to select the trial where the partner is most likely to do the opposite as what she did before.

Just like the ideal search model, all three alternative models make choices in a stochastic manner.

2.5. Model fitting

All computational models are built on top of the ideal observer model, which must be equipped with a prior. I chose the shape of the prior on the basis of empirical data about the way people play the WTT (Sznycer, Lim, Tooby & Cosmides, unpublished data)[6]. The empirical distribution of WTRs shows a sharp discontinuity at WTR = 0; see Fig. 2. Their distribution is well-approximated by a skewed Laplacian distribution, which has a peak at WTR = 0 and declines faster on the negative tail than on the positive tail (i.e. very few people have negative WTRs).

Therefore I assume that people's priors take the shape of a skewed Laplacian distribution with a peak at 0. This family of distributions has two other parameters, namely skew and dispersion.

In sum, each model has three free parameters that need to be fit to the human data in the main task: the β parameter (which determines the stochasticity of the model's choices), and the skew and dispersion of the ideal observer's prior.

For each model, I fit these three parameters to the human data by finding the parameter values that maximized the log-likelihood of participants' choices according to the model. Best-fitting values for the ideal search model were skew = .24, dispersion = .38, β = 2.16. Fig. 3 plots the corresponding prior distribution (skewed Laplacian with location = 0, skew = .24, dispersion = .38).

[5] https://osf.io/jtavm

[6] I thank Daniel Sznycer for sharing these data.

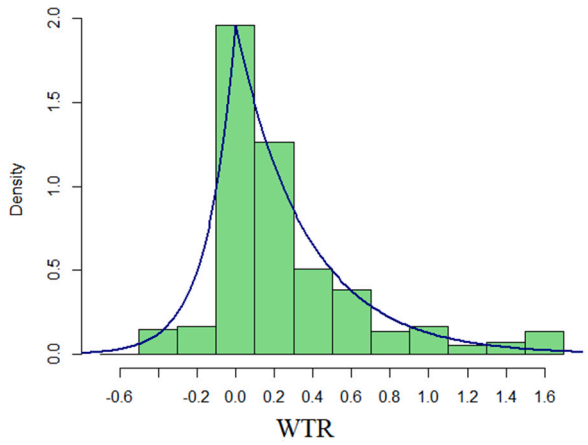

Fig. 2. Distribution of participants' WTRs in Sznycer et al. (unpublished data). In blue, the best-fitting probability density function.

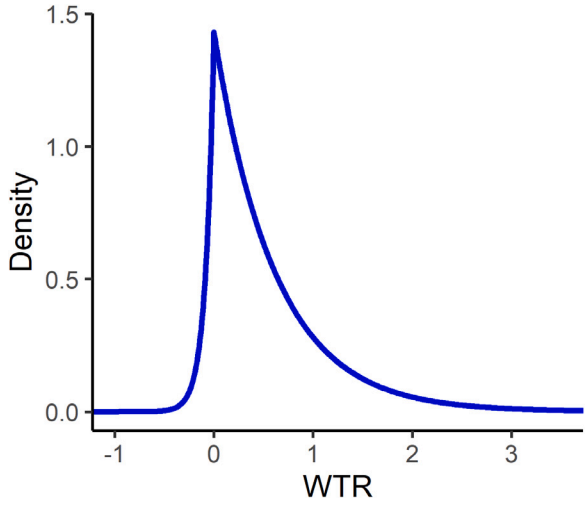

Fig. 3. Prior belief of the ideal observer about the partner's WTR. Mean = .38, Standard Deviation = .44.

Finally, to limit the number of free parameters, I set the value of σ_ϕ, the amount of noise in Alice's decisions, to $\sigma_\phi = .16$. This is the amount of noise that is most consistent with the pattern of choices made by people playing the WTT as dictators (in the same data by Sznycer et al. unpublished. See appendix for details).

The data, and the R code for the computational model, data analysis and figures are available at the Open Science Framework.[7]

3. Study 1

The current experiment is designed to test whether human data selection in a simple task is well-described by the ideal search model proposed above. Participants were paired with a sham partner, playing the WTT as recipients with the partner playing as dictator. Participants were shown the decision of their partner in one trial of the WTT, after which they were shown pairs of trials, for which they could see the payoffs involved but not the partner's decision. For each pair of trials, I asked participants for which trial they most would want to know the decision made by their partner. I predicted that participants would show more curiosity toward the trials that had the highest expected information value regarding the partner's WTR, as measured by the ideal search model.

Note that, in contrast to most experiments on human data selection, the current task had no explicit 'correct' answer. I simply asked participants which trial they would most want to see, did not instruct them to maximize their information intake, and did not incentivize their choices. Therefore, to a certain extent this task measures 'spontaneous' curiosity.

3.1. Participants

I recruited 216 participants from the undergraduate psychology participant pool at the University of California Santa Barbara, who participated in exchange for course credit (the stopping rule for participant recruitment was to stop after the day I reached 200 participants or more). I excluded from analysis 71 participants who failed either an attention check ($N = 55$) and/or a comprehension question (N = 22), leaving a total of 145 participants (95 female, mean age : 18.9, sd : 1.40).[8]

3.2. Procedure

Participants completed the study on a desktop computer while seated in a semi-private cubicle. They were first given a description of the WTT, and played a few rounds of a pretend version of the task in the role of dictator, in order to get familiarized with the task.

In the main phase of the study, participants were asked to imagine that they were playing the WTT in the role of the recipient. They were shown information about the choices faced by a computer-generated partner playing as dictator, and were asked to imagine that this partner was one of their acquaintances. Participants were aware that their partner was computer-generated, and no deception was involved in the task. All monetary payoffs were fictitious.

To manipulate participants' prior beliefs about their partner, we first showed them the outcome of one decision made by their partner. This decision was manipulated between-subjects: half of participants saw their partner make a selfish decision (allocate $30 to themselves instead of allocating $30 to the participant), while the other half saw their partner make a generous decision (allocate $30 to the participant instead of $10 to themselves). These decisions were designed so that they would yield enough information to shift the belief of the ideal observer when observed, but not so much information that they would virtually eliminate the usefulness of subsequent information. For instance, observing Alice giving $30 to Bob instead of taking $10 suggests that she is relatively generous, but does not tell us exactly how generous she is.

Henceforth I refer to the first condition as the 'Take' condition and the second condition as the 'Give' condition. To increase the likelihood that participants would process this initial information, I asked them to rate how grateful and how angry they were at their partners after observing the decision (on two 1–7 likert scales, with 1: "not at all" and 7: "very grateful"/"very angry").

Then, in the critical phase of the experiment, participants were shown fifteen pairs of WTT trials on which that same partner had made decisions. They were shown the payoffs involved in each trial (i.e. the values of $\pi_{partner}$ and $\pi_{participant}$ for each trial) but not the decision that their partner had made. Trials were created by using values for $\pi_{partner}$ drawn from the set {-$15, $3, $21, $39, $57, $75}; $\pi_{participant}$ was

[7] https://osf.io/jtavm

[8] The relatively high number of participants failing the attention check might be due to the fact that it required participants to inhibit a natural impulse (they were asked to answer "green" to the question "what is the color of the ocean?"). For both studies, analysis on all participants yields similar results as after exclusions (see the Supplementary Information at https://osf.io/jtavm).

always $30. I created one pair of trials for each possible combination of payoffs to the partner, subject to the constraint that the two trials within a pair could not have the same value of $\pi_{partner}$, resulting in fifteen different pairs of trials. Trial pairs were presented in randomized order.

For each pair of trials, I asked participants for which trial they would most like to see the decision made by their partner, using a binary question. Each pair of trials was presented on a separate page of the computer-based survey. For each pair of trials, the order in which the trials were displayed on the page was counterbalanced across participants. On the top of each page, I also reminded participants of the first decision made by their partner. I did not give feedback to participants: giving them more information about their partner's decisions would have changed their estimates of the partner's WTR, weakening experimental control.

Additionally, participants completed two unrelated tasks (whose results I report elsewhere; Quillien, 2021). The first task was a prediction task where participants were asked to predict the behavior of other players in the WTT. Half of participants completed the prediction task before the data selection task, while the other half completed that task after the data selection task. The second task was an emotion rating task, which probed participants' anger and gratitude toward 10 different partners making one decision each. All participants completed the emotion ratings task after the data selection task.

Then participants were asked a few demographic questions (age, gender, ethnicity, whether English was their native language, and whether they had taken a college-level statistics or probability class) and were thanked for their participation.

4. Results

Fig. 4 displays the average proportion of participants making a given choice for every pair of trials. Fig. 5 plots the same information for the choices made by the ideal search model.

The first observation one can make is that participants did not choose at random: for the vast majority of trial pairs, people's choices significantly differ from the chance level of 50%. Second, people seemed to make choices that intuitively feel informative. For instance, about 75% of people in the 'Give' condition selected the trial with {$21 for partner, $30 for participant} as more interesting than the trial with {$3 for partner, $30 for participant}. That is, people who have prior information suggesting that their partner is relatively generous seem relatively uninterested by a trial for which they should be confident that the partner will Give. Third, the red and blue lines are not exactly superimposed, suggesting that the between-subjects manipulation made a difference to participants' choices.

Do participants' selections reflect the expected information content of the trials?

Yes. People tended to select the trials with the highest expected information value.

For each trial pair, I computed the average proportion of participants making a given choice, and the probability that the ideal search model would make that same choice. The item-level correlation between people's average choices and the choices made by the ideal search model was r(28) = .878, p <.001. Fig. 6 depicts the correlation between ideal search model and participant average choices, broken down by condition.

Do participants with different prior information select different data?

Yes. The information content of a given trial depends on the prior beliefs of an observer; therefore the ideal search model selects different trials depending on whether it has previously observed the partner Give or Take. Fig. 7 shows that the ideal search model will select trials with a higher value of $\pi_{partner}$ if it has seen its partner make a generous decision before. Human choices followed the same pattern: participants in the "Give" condition selected trials with a higher value of $\pi_{partner}$ than participants in the "Take" condition; b = −4.5, p = .02; (linear mixed model with random intercepts, participants as random effect).

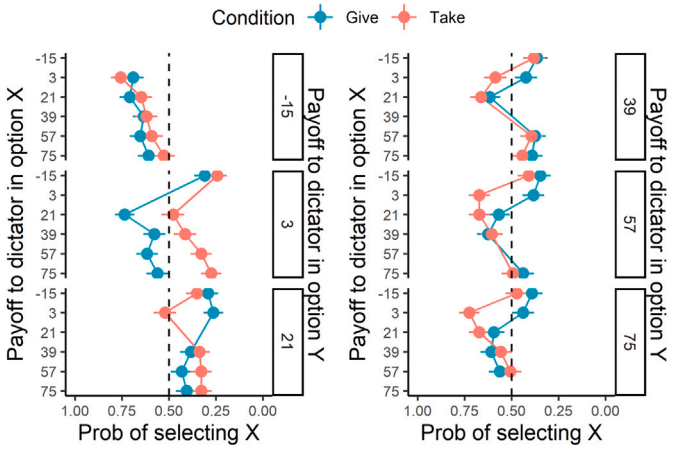

Fig. 4. Average proportion of participants selecting a trial, for every possible pair of trials, Study 1. Payoffs are in USD, and potential payoff to recipient was $30 in every trial. For instance, about 75% of people in the 'Give' condition selected the trial with {$21 for partner, $30 for participant} as more interesting than the trial with {$3 for partner, $30 for participant}. Error bars represent the standard error of the mean. Note that each trial pair is plotted twice. For instance, the data point for $\{\pi_{partner} = 3$ vs $\pi_{partner} = 21\}$ represents the same data as the data point for $\{\pi_{partner} = 21$ vs $\pi_{partner} = 3\}$.

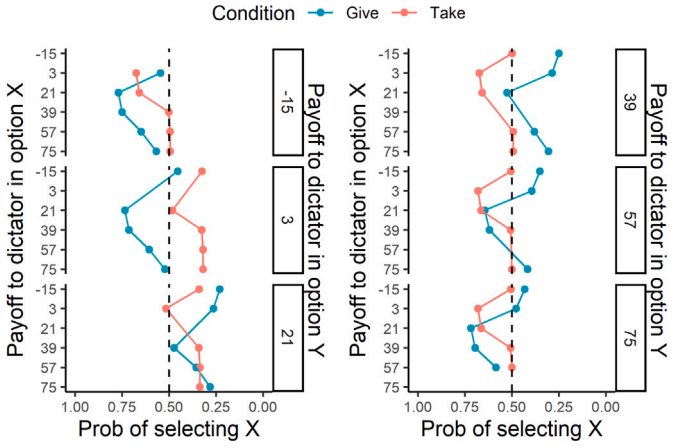

Fig. 5. Probability that the stochastic ideal search model selects a trial, for every possible pair of trials. Payoffs are in USD, and potential payoff to recipient was $30 in every trial. For instance, in the 'Give' condition the ideal search model selected the trial with {$21 for partner, $30 for participants} over the trial with {$3 for partner, $30 for participant} with probability .75. Note that each trial pair is plotted twice. For instance, the data point for $\{\pi_{partner} = 3$ vs $\pi_{partner} = 21\}$ represents the same data as the data point for $\{\pi_{partner} = 21$ vs $\pi_{partner} = 3\}$.

Did participants simply select trials with the highest (or the lowest) value of $\pi_{partner}$?

No. The choices of the ideal search model followed an inverted-U curve, and people's choices followed a similar pattern (see Fig. 7). To test for the statistical significance of this inverted-U curve pattern in the human data, I performed two-lines tests (Simonsohn, 2018).

For participants in the Take condition, in the interval between $\pi_{partner} = -15$ and $\pi_{partner} = 3$, the value of $\pi_{partner}$ in a trial was a positive predictor of the probability of selecting that trial; b = .10, p <.001 (multilevel logistic regression with random slopes and random intercepts, and participants as random effects). In the interval between $\pi_{partner} = 3$ and $\pi_{partner} = 75$, it was a negative predictor, b = -.02, p <.001.

For participants in the Give condition, in the interval between $\pi_{partner} = -15$ and $\pi_{partner} = 21$, the value of $\pi_{partner}$ in a trial was a

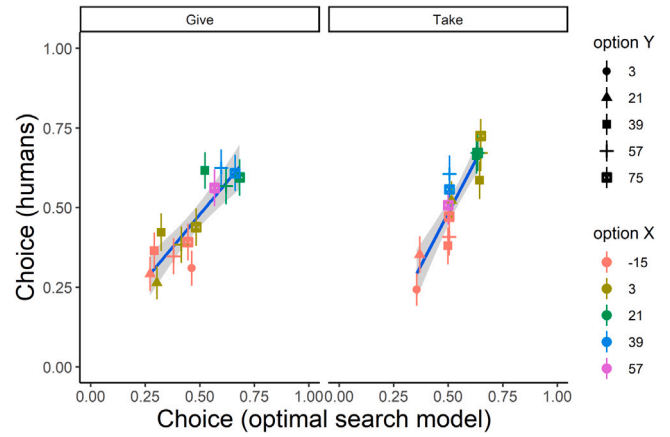

Fig. 6. Relation between the choice probability of the optimal search model and the choice probability of human participants, presented separately for participants in the Give and the Take condition, Study 1. Each point represents one pair of trials. Error bars represent the standard error of the mean. Higher values correspond to a higher probability of choosing option X.

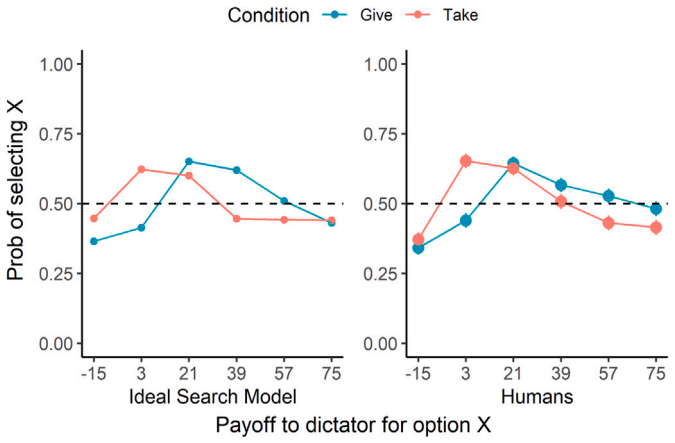

Fig. 7. Probability of selecting a trial, as a function of potential payoff to dictator for that trial, for the ideal search model (left) and human participants (right), Study 1. This graph collapses over all other potential values of $\pi_{partner}$ for the other trial in the pair. Error bars represent the standard error of the mean.

positive predictor of the probability of selecting that trial; b = .05, p <.001 (multilevel logistic regression with random slopes and random intercepts, and participants as random effects). In the interval between $\pi_{partner}$ = 21 and $\pi_{partner}$ = 75, it was a negative predictor, b = -.02, p = .02.

In sum, participants were not consistently attracted to trials with extreme values of the potential payoff to the dictator. Instead, their choices followed the pattern of choices of the ideal search model.

Do alternative models account for the data?

The selections of the 'ideal search without updating' model are shown in Fig. 8. The item-level correlation between people's choices and the choices made by the model was r(28) = .812, p <.001; slightly lower than the r = .878 achieved by the ideal search model.

The item-level correlation between people's choices and the choices made by the falsification model was r(28) = .525, p = .003; for the confirmation model, this correlation was negative, r(28) = -.100, p = .60.

I also computed the fits of the different models to the human data by computing, for each choice made by one participant, the log-likelihood of this choice according to the model, and then summing all these log-likelihoods across all choices and all participants. Table 1 shows the log-likelihood, and the item-level correlation, for each model.

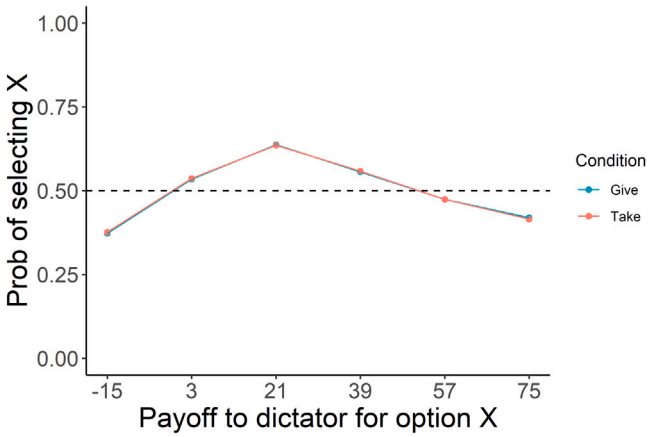

Fig. 8. Probability that the ideal-search-without-updating model selects a trial, as a function of potential payoff to dictator for that trial. This graph collapses over all other potential values of $\pi_{partner}$ for the other trial in the pair. The model makes the same predictions for both conditions because it does not update its belief on the basis of new observations.

Table 1
Log-likelihood, and Pearson's correlation coefficient (r) for the fit of each search model to the human data, Study 1.

Model	Log-likelihood	Pearson's r
Ideal search	−1424	.878***
Ideal search no updating	−1435	.812***
Falsification	−1468	.525**
Confirmation	−1488	−.100 (n.s)

Higher log-likelihoods (i.e. closer to 0) indicate better fit. ***: p < .001; **: p < .01; n.s : p > .05.

5. Study 2

Results of study 1 suggest that people spontaneously select the evidence that contains the most potential information about the WTR of their partner.

Participants' choices were not driven solely by the specific payoffs in each trial. Given a choice involving the same pair of trials, participants tended to select a different trial depending on whether their partner had acted selfishly or generously before. As predicted by an optimal mathematical model of data selection, participants' selections were shaped by an interaction between the properties of the trials and the prior information that participants had available.

A potential deflationary explanation for the current results is that participants were not trying to infer the WTR of their partner. Instead, they were simply curious about the payoff they would get, and selected the option in each pair for which the outcome was most uncertain.[9]

In the current data selection task, the trials that contain the most expected information about the partner's WTR are also the trials in which the partner's decision is least predictable (technically, the trials that have highest information entropy). This raises the question, are the trials that participants find more interesting simply ones for which the outcome is most uncertain? That is, maybe people were curious about the outcome of the trial (whether they gained money or not), rather than the WTR of their partner. I will call this interpretation the "outcome-oriented" account. This account has some prior plausibility, given that people sometimes use their uncertainty about the outcome of an observation as an (imperfect) proxy for its information value (Markant & Gureckis, 2014).

[9] Note that this account is not entirely deflationary: computing outcome uncertainty still requires a causal model of how others make welfare trade-offs, and the ability to make approximately Bayesian computations.

In study 2, I attempt to rule out this interpretation. In order to de-confound information entropy and information value, I introduce additional pairs of trials to choose from, where participants are asked to assume that the outcome of one trial is decided by a person (as before), but the outcome of the other trial is decided by a coin flip by the computer. I will refer to these as hybrid pairs. In hybrid pairs, the coin flip option has maximum information entropy (its outcome is completely unpredictable), but it contains no information about the partner's WTR. The WTR inference account predicts that people should prefer to look at trials where their partner, rather than the computer, is determining the outcome, despite the fact that the outcomes of the computer-determined trials are more uncertain. The outcome-oriented account predicts that people will be more likely to choose the maximally uncertain coin-flip trial.

Study 2 also attempts a direct replication of the results of study 1, in a different and larger sample. While Study 1 was conducted with undergraduate students, in Study 2 I collect data from a more diverse sample of participants, recruited on the online recruitment platform Prolific.

5.1. Participants

I recruited 300 US residents from Prolific, an online platform. I excluded from analysis 107 participants who failed either an attention check ($N = 60$) and/or one of three comprehension questions (Ns = 26, 38, 32), leaving a total of 193 participants (99 male, 91 female, 3 other, mean age: 34.1, sd: 12.8).

5.2. Procedure

Study 2 was identical to Study 1, with the following exceptions. First, I omitted the unrelated prediction and emotion tasks. Second, in the data selection task, in addition to the 15 pairs of WTT trials where both decisions were made by the participant's partner, there were 6 'hybrid' pairs of trials for which the outcome of one trial was determined by the computer, and the outcome of the other trial was determined by the participant's partner. I told participants that in a WTT trial whose outcome is determined by the computer, the computer simply chooses randomly whether to allocate money to the participant or the partner. Two comprehension questions in the instruction phase of the study probed whether participants understood that there was a 50% probability of either player getting money in such trials (participants failing any of these questions were excluded from analysis). In addition, in computer-determined trials, a picture of a coin flip on the participant's screen served as a reminder of the probabilistic nature of the computer's "decision". The values of $\pi_{partner}$ (in USD) for each pair of trials were the following (C: computer, P: partner; each bracket represents one pair): {C:-15, P:21}, {C:3, P:39}, {C:21, P:57}, {C:39, P:75}, {C:57, P:-15}, {C:75, P:3}. The value of $\pi_{participant}$ was always $30. According to the outcome-oriented account, participants should always select the computer-determined trial, regardless of the content of a trial pair. Hybrid trial pairs were randomly interspersed among normal trial pairs.

The ideal search model was identical to the one used in study 1, except that its free parameters (for the prior, and the softmax choice selection function) were fit to the data selection choices of participants in the current study.

5.3. Results

I first discuss whether results of Study 1 are replicated, looking only at participants' selections for normal trial pairs. Then I discuss results for the new hybrid trial pairs separately.

Do participants select data with high information content?

Yes. The item-level correlation between people's choices and the choices made by the ideal search model was r(28) = .841, p <.001.

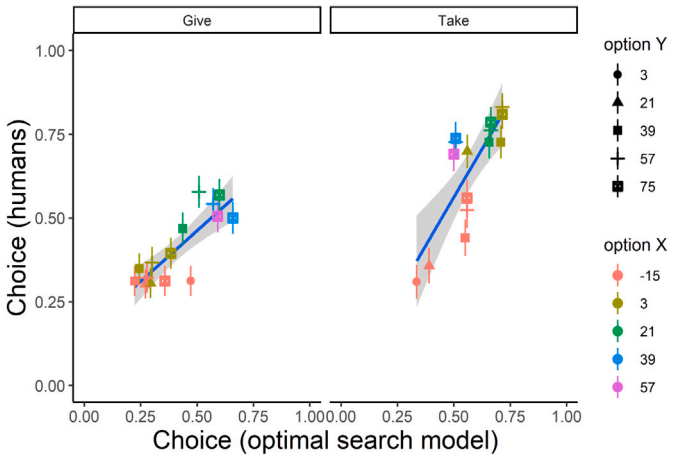

Fig. 9. Relation between the choice probability of the optimal search model and the choice probability of human participants, displayed separately for participants in the Give and the Take condition, Study 2. Each point represents one pair of trials. Error bars represent the standard error of the mean. Higher values correspond to a higher probability of choosing option X.

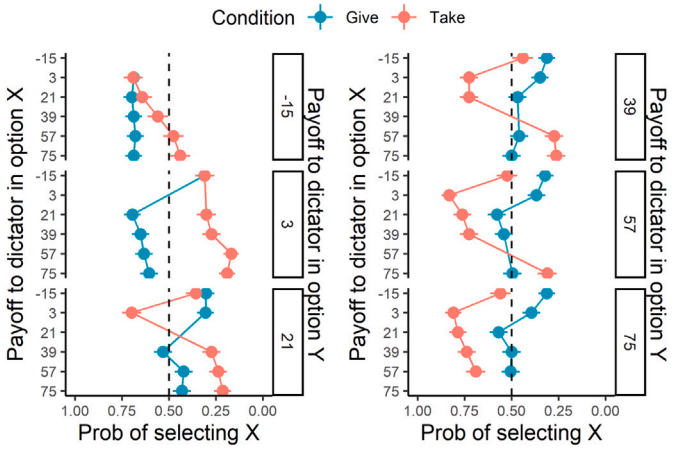

Fig. 10. Proportion of human participants selecting a trial, for every possible pair of trials, Study 2. Payoffs are in USD, and potential payoff to recipient was $30 in every trial. For instance, in the 'Give' condition participants selected the trial with {$21 for partner, $30 for participants} over the trial with {$3 for partner, $30 for participant} 75% of the time. Note that each choice is plotted twice. For instance, the data point for {$\pi_{partner} = 3$ vs $\pi_{partner} = 21$} is the same as the data point for {$\pi_{partner} = 21$ vs $\pi_{partner} = 3$}.

Fig. 9 depicts the correlation between ideal search model and participant average choices, broken down by condition.

Fig. 10 displays the average proportion of participants making a given choice for every trial pair. Fig. 11 plots the same information for the choices made by the ideal search model.

Do participants with different prior information select different data?

Yes. The information content of a given trial depends on the prior beliefs of an observer; therefore the ideal search model selects different trials depending on whether it has previously observed the partner Give or Take. Fig. 12 shows that the ideal search model will select trials with a higher value of $\pi_{partner}$ if it has seen its partner make a generous decision before. Human choices followed the same pattern: participants in the "Give" condition selected trials with a higher value of $\pi_{partner}$ than participants in the "Take" condition; b = −10.6, p <.001; (linear mixed model with random intercepts, participants as random effect).

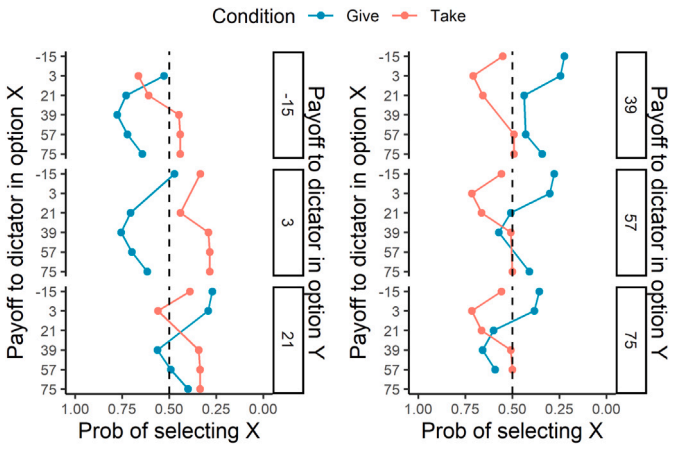

Fig. 11. Probability that the stochastic ideal search model selects a trial, for every possible pair of trials. Payoffs are in USD, and potential payoff to recipient was $30 in every trial. For instance, in the 'Give' condition the ideal search model selected the trial with {$21 for partner, $30 for participants} over the trial with {$3 for partner, $30 for participant} with probability .75. Note that each choice is plotted twice. For instance, the data point for $\{\pi_{partner} = 3 \text{ vs } \pi_{partner} = 21\}$ is the same as the data point for $\{\pi_{partner} = 21 \text{ vs } \pi_{partner} = 3\}$.

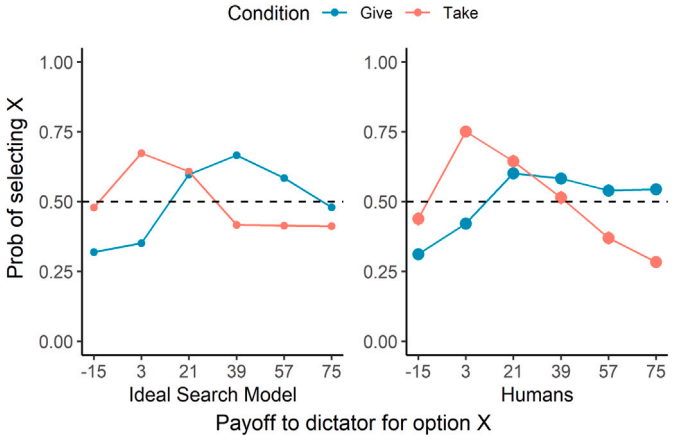

Fig. 12. Probability of selecting a trial, as a function of potential payoff to dictator for that trial, for the ideal search model (left) and human participants (right), Study 2. This graph collapses over all other potential values of $\pi_{partner}$ for the other trial in the pair. Error bars represent the standard error of the mean.

Did participants simply select trials with the highest (or the lowest) value of $\pi_{partner}$?

No. The choices of the ideal search model followed an inverted-U curve, and people's choices followed a similar pattern (Fig. 12). To test for the statistical significance of this inverted-U curve pattern in the human data, I performed two-lines tests (Simonsohn, 2018).

For participants in the Take condition, in the interval between $\pi_{partner} = -15$ and $\pi_{partner} = 3$, the value of $\pi_{partner}$ in a trial was a positive predictor of the probability of selecting that trial; b = .10, p <.001 (multilevel logistic regression with random slopes and random intercepts, and participants as random effects). In the interval between $\pi_{partner} = 3$ and $\pi_{partner} = 75$, it was a negative predictor, b = -.04, p <.001.

For participants in the Give condition, in the interval between $\pi_{partner} = -15$ to $\pi_{partner} = 21$, the value of $\pi_{partner}$ in a trial was a positive predictor of the probability of selecting that trial; b = .05, p <.001 (multilevel logistic regression with random slopes and random intercepts, and participants as random effects). However, in the interval between $\pi_{partner} = 21$ and $\pi_{partner} = 75$, $\pi_{partner}$ for a trial had no effect on the likelihood of selecting that trial, b = .00, p = .88.

Table 2
Log-likelihood, and Pearson's correlation coefficient (r) for the fit of each search model to the human data, Study 2.

Model	Log-likelihood	Pearson's r
Ideal search	−1880	.841***
Ideal search no updating	−1932	.635***
Falsification	−1952	.520**
Confirmation	−2004	−.413*

Higher log-likelihoods (i.e. closer to 0) indicate better fit. ***: p < .001; **: p < .01; *: p < .05; n.s : p > .05.

In sum, participants were not consistently attracted to trials with extreme values of the potential payoff to the dictator. Instead, participants who had seen their partner make a selfish decision had the same pattern of choices as the ideal search model. Participants who had seen their partner make a generous decision had a pattern of choices close the ideal search model, except that at high values of $\pi_{partner}$ the effect of $\pi_{partner}$ was flat instead of decreasing.

Do alternative models account for the data?

The item-level correlation between people's choices and the choices made by the ideal-search-without-updating model was r(28) = .635, p <.001; lower than the r = .841 achieved by the ideal search model.

The item-level correlation between people's choices and the choices made by the falsification model was r(28) = .520, p = .003; for the confirmation model, this correlation was negative, r(28) = -.414, p = .02.

I also computed the fits of the different models to the human data by computing, for each choice made by one participant, the log-likelihood of this choice according to the model, and then summing all these log-likelihoods across all choices and all participants. Table 2 shows the log-likelihood, and the item-level correlation, for each model.

Were participants curious about their immediate payoffs, or about their partner's WTR?

According to the outcome-oriented account, when participants can request to observe either a computer-generated or a partner-generated decision, they should always be biased toward the computer-determined decision, regardless of the content of a trial pair. Participants actually showed the reverse bias: on average, across all hybrid trials, they chose to observe their partner's decision 57% of the time. This was significantly larger than the chance level of 50%, p <.001, as indicated by the intercept of a multilevel logistic regression with random intercepts, participant as random effect, and no independent variable.

Fig. 13 shows participants' choices in more detail. For trial pairs in which observing the partner's decision has very low expected information value (for instance, when it would cost a selfish partner $75 to give the participant $30), participants tended to choose randomly, even though the computer-determined trials had much greater information entropy. When the partner's decision had large expected information value, participants were strongly inclined to observe it, doing so about 70% of the time.

In sum, the results of the hybrid trials are consistent with the WTR-inference interpretation over the outcome-oriented account.

5.4. Study 2 discussion

Study 2 replicates the main results of Study 1: participants tended to be curious about the trials that would reveal the most information about their partner's WTR. In addition, it shows that this pattern does not arise because participants are simply interested in the decisions' outcomes per se. Instead, their selections are the output of a psychology designed to extract information about a causally deep property of the social world: someone's valuation of your welfare.

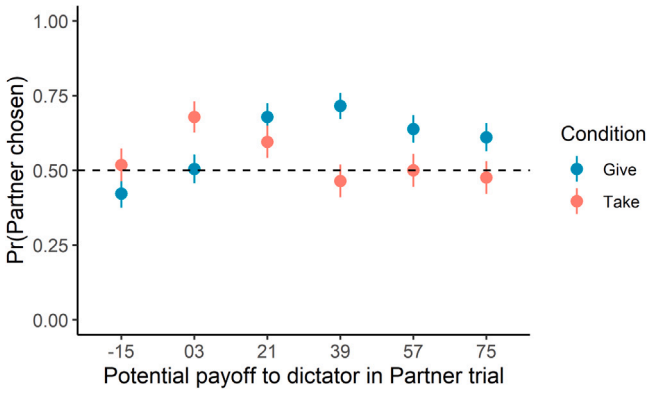

Fig. 13. Proportion of participants who selected the partner-determined trial instead of the computer-determined trial, as a function of the potential payoff to the partner. Error bars represent the standard error of the mean.

6. General discussion

According to a growing body of research, representations of how much someone values our welfare play a fundamental role in social cognition. This view implies that people have cognitive mechanisms that allow them to efficiently construct these representations. Here I tested this prediction in the context of information-gathering.

In a simple data selection task, participants tended to request the data that was most informative about how much someone valued their welfare. Participants did so even though I did not give them explicit criteria for how to make their choices: I simply asked them which decision they would most like to observe. I did not give participants any information that might have suggested that they would need to infer the weight that their partner put on their welfare.[10] In sum, people spontaneously showed interest for data that were most informative about social valuation.

Below I discuss these results in the context of the psychology of information-gathering and theories of social cognition, and discuss directions for future research.

6.1. The psychology of information-gathering

The current study adds to a body of research that shows that people are often able to select information in near-optimal ways in a wide range of domains (Najemnik & Geisler, 2005; Nelson, 2005; Oaksford & Chater, 1994). I show that in an evolutionarily important social domain, people can select queries according to their expected information value.

The study also differed from most studies on data selection, in that it involved pre-existing domain-specific knowledge. Most studies of information search give participants explicit information about the goal of the task and its information structure. For instance, in a typical task (e.g. Nelson et al., 2010) participants first learn to sort individual items into two categories by observing many items for which they can see the features and the category label. Then in a test phase they have to categorize new items, and can request which features of the items they want to see. In such studies, the knowledge learned in the observation phase allows the participant to subsequently compute the information value of each feature in the test phase. Therefore these tasks do not require any pre-existing domain-specific knowledge.

By contrast, human cognition heavily relies on domain-specific knowledge, organized into intuitive theories (Gerstenberg & Tenenbaum, 2017; Hirschfeld & Gelman, 1994). In the current task, participants only observed one decision made by their partner before the main task — therefore they did not have the opportunity to learn about the way that people typically behave in money allocation games. Instead they had to use their pre-existing knowledge about how people make welfare tradeoffs. As such, the present results show that people are able to spontaneously mobilize their domain-specific causal knowledge in order to guide information-gathering.

Participants made choices that were in line with a computational-level analysis of the task (Anderson, 1990; Marr, 1982). But the current data do not speak to the algorithms that they used to make their choices. Participants may not have performed the same complex computations that are prescribed by the ideal search model (i.e. computing the expected information value of a query across all possible outcomes, by computing the rational probabilistic inferences warranted by each potential outcome). There has been a growing interest in discovering the heuristic strategies that people use to efficiently select data in a resource-rational way (Coenen et al., 2019; Liefgreen, Pilditch & Lagnado, 2020). As an example, below is a speculative hypothesis for the mechanism underlying participants' performance.

Asking about the events whose outcome you are most uncertain about is often a good strategy to maximize expected information gain (in the current task, it is a good strategy in trials that do not involve coin flips). You are most uncertain about what your partner will do when you estimate that the decision she faces is close to her point of indifference: for instance if you think her WTR is around .4, and the payoffs are such that she would give if and only if her WTR is above .4, then from your standpoint your partner is equally likely to Give and Take. Therefore, one approximate strategy to maximize outcome uncertainty is to ask to observe trials that you think are close to your partner's point of indifference. If your best estimate of your partner's WTR is .4, you can (approximately) maximize your outcome uncertainty by asking to observe a decision for which the ratio of payoffs is close to .4. Maybe participants used something close to this strategy in the trials where outcome uncertainty is a good proxy for expected information gain (i.e. in trials that do not involve coin flips).

Future research could use a more complex version of the current task, with the goal of observing situations where participants make systematic deviations from normative principles. An analysis of their patterns of errors might yield valuable information about the strategies they use.

6.2. Social cognition

Humans are an extraordinarily cooperative species. Evolutionary biologists and cognitive scientists have suggested that our rich social life is made possible by a set of emotional systems that motivate us to monitor and regulate the way that other people treat us (Cosmides & Tooby, 2000; Pinker, 1997; Tooby & Cosmides, 2008; Trivers, 1971). There is indeed a lot of evidence that people care about how much others value them (Delton & Robertson, 2012; Eisenbruch & Krasnow, 2022; Eisenbruch & Roney, 2017; Fiske et al., 2007; Lim, 2012; Sell, 2005), and are motivated to find information about it (Brambilla et al., 2011; Wojciszke et al., 1998). But relatively little is known about how the mind constructs representations of social valuation. Here I find that, at least in a simple task, people can efficiently gather the information relevant to building these representations.

This finding adds support to the idea that causal inferences about the mental states and traits of others play a key role in social cognition (Ajzen & Fishbein, 1975; Fiske et al., 2007; Heider, 1958; Sell et al.,

[10] In study 1, participants also completed a prediction task — one might argue that this primed them to look for WTR-relevant information during the data selection task. I tested this possibility by using the fact that half of the participants completed the prediction task before the data selection task, and the other half completed it after the data-selection task. A multilevel logistic regression finds that the relationship between ideal search model predictions and participant choices in the data selection task is not moderated by task order (interaction: $p = .44$). And the prediction task was completely absent in study 2, which replicated study 1's results.

2017). From a functional perspective, making such causal inferences is useful because it allows us to more accurately predict and influence a person's behavior in the future. By contrast, simple heuristic rules, that respond to what a partner did but not why they did it, might yield suboptimal outcomes. For example if our partner failed to help us in one particular instance where helping would have been prohibitively costly for her, we should not necessarily conclude that she will not help us under more normal conditions (Lim, 2012; Qi & Vul, 2020, 2022; Tooby et al., 2008).

To succeed in the current task, participants needed to possess a domain-specific causal model of how agents make welfare trade-offs. What is the origin of this causal model?

Humans have a set of mechanisms that allow them to reason about the minds of others. Recent models of commonsense psychology hold that we are "naïve econometricians" when reasoning about the behavior of other agents. People hold that agents have preferences, and they explain behavior with the assumption that agents behave in a rational way toward realizing their preferences (Baker et al., 2017; Jara-Ettinger et al., 2016; Jara-Ettinger, Schulz, & Tenenbaum, 2020; Jern et al., 2017; Lucas et al., 2014; Quillien & German, 2021).

The idea that people represent the welfare-tradeoff parameters of others fits well within this framework (see Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013; Kleiman-Weiner, Saxe, & Tenenbaum, 2017; Powell, 2022; Qi & Vul, 2022; Ullman et al., 2009). People may represent the fact that Alice values the welfare of Bob half as much as her own with the same mechanisms they use to represent the fact that Alice likes oranges half as much as apples. These mechanisms allow people to make rational inferences about the preferences of others after observing their decisions (Baker et al., 2017; Jara-Ettinger et al., 2016; Lucas et al., 2014), and could therefore allow people to infer the WTPs of an agent on the basis of her actions. As such, the current results contribute new evidence for the theory that people are naive econometricians. While previous work has shown that people can make rational inferences about the preferences of other agents (Baker et al., 2017; Jern et al., 2017), here I find that people can select the data that are most informative for making these inferences.

It is also likely that humans have strong reliably-developing inductive biases that help them quickly build a good model of how people make welfare-tradeoffs. Representations of social valuation are an input to motivational systems: this fact suggests that the concept of welfare tradeoff might be part of our species' standard equipment. People everywhere get angry in response to cues that they are under-valued (Sell et al., 2017), feel pride in response to traits or achievements that make others more likely to value them (Sznycer et al., 2017; Sznycer, Xygalatas, Alami et al., 2018), and feel shame in response to traits and events that make others less likely to value them (Sznycer et al., 2016b; Sznycer, Xygalatas, Agey et al., 2018). Social valuation inferences also strongly motivate us to recalibrate our own WTPs (Lim, 2012; Smith et al., 2017) toward others. The fact that natural selection was able to design these motivational systems suggests that an understanding of social valuation reliably develops in humans (Delton & Sell, 2014; Tooby et al., 2005, 2008).

6.3. Limitations and future directions

In this study I used a very simple data selection task. The different dilemmas that participants could choose to observe only varied along one parameter (the opportunity cost of giving). Would people also perform well in tasks that vary other parameters, or vary several parameters at once? For example, whether an action was made intentionally moderates how much the action reveals about the actor's disposition toward you. How well do people factor this fact when they compute the information content of an action?

Someone can value your welfare because she is a nice person in general, or because she values you in particular. This is an important distinction, as people are especially concerned about whether others value them in particular (Krasnow, Delton, Cosmides, & Tooby, 2016; Lukaszewski & Roney, 2010). The current task conflates these two sources of variation in welfare-tradeoff behavior. Therefore, it is unclear if participants were interested in WTP-relevant information because it would reveal their partner's general level of generosity, or their partner's valuation of them in particular. A natural extension of the current task would be to introduce trials where our partner makes decisions involving a third party. A functional perspective predicts that people should be more curious about trials involving themselves than trials involving third parties, but only up to a point. People expect that the way someone treats a third-party contains some information about how they might treat you (Krasnow et al., 2016). As such, participants might sometimes consider that a third-party trial contains more information than a trial involving themselves, for instance if the outcome of the former trial is uncertain while the latter's is already obvious. The current computational model could be extended to model this task, for example by adding assumptions about how people generalize on the basis of third-party valuation.

More generally, inferring and predicting how other people make welfare-tradeoffs is a complex task, a full model of which is beyond the scope of the current work. Many different factors shape human welfare-tradeoff psychology: I might help you because we are siblings, because you helped me in the past, because we belong to the same group, because of shared social norms, etc (Tooby et al., 2008). In order for you to predict how I will behave in the future on the basis of what I did, you need to know (or guess) *why* I helped you. Put more formally, a full model of welfare-tradeoff inference would involve hierarchical Bayesian inference, where one not only infers WTPs but their underlying causes (see Kleiman-Weiner et al., 2017). The psychology of welfare-tradeoffs is also connected to other parts of moral cognition, and people can make inferences about someone's proneness to harm even on the basis of non-harmful norm violations (Chakroff, Russell, Piazza, & Young, 2017).

By design, the queries available to participants could only give them information about their partner's WTPs — as opposed to information about their partner's skill or competence, for instance. That is, the task was designed to investigate people's ability to select the most informative data, but it was not designed to study whether people are more interested in social valuation than by other facts about a social partner. Existing studies suggest that people do prioritize social valuation information over information about other traits in their data selection (Brambilla et al., 2011; Wojciszke et al., 1998). Future research might use formal measures of information value to more precisely quantify to what extent people prioritize information about social valuation over other traits.

Here I measured the value of information with methods from probability theory, which are agnostic about the domain that the information is about. A complete account of information search in the social domain would benefit from a more domain-specific task analysis. Such a task analysis would take into account the utility of information with respect to people's goals. For instance, someone who wants to choose who to cooperate with might try to gather information allowing him to choose between two potential partners he already knows to be generous, and disregard information about people he already discarded as potential partners. In other contexts, information that reveals that someone is a cheater might be more valuable than information revealing that someone is extremely generous. Indeed, people can be sensitive to such asymmetries when they select which questions to ask in order to form an impression of a new person (Brambilla et al., 2011). Some people might also prefer seeking information that they think will make them feel better, even when this conflicts with the goal of collecting practically useful information (Kelly & Sharot, 2021).

Finally, while researchers sometimes study social cognition by having participants interact with actual people (or persuading them that they are), in the current study participants were aware that they were

not actually playing with another human. However, since the current work investigates inference (rather than affect or motivation), there are no strong reasons to expect that this design choice influenced the results (additionally, participants in economic games who know they are playing with fake partners do not behave much differently than other participants; (Krasnow, Howard, & Eisenbruch, 2020).

7. Conclusion

Evolutionary theory can suggest hypotheses about the kinds of representations that the human mind is designed to construct. Testing these hypotheses using the tools of Bayesian cognitive science is a promising way to reverse-engineer the structure of cognitive mechanisms (Pietraszewski & Wertz, 2011).

Using this approach, I find that people seem to be rationally curious about social valuation: they spontaneously tailor their information search toward the data that is potentially most revealing about how much someone values them. This finding suggests that humans can rely on pre-existing domain knowledge to make near-optimal queries in data selection tasks. It also provides evidence in favor of the view that the human mind houses cognitive machinery that models the welfare-tradeoff behavior of others.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All code and data have been made available at the Open Science Framework at https://osf.io/jtavm.

Acknowledgments

This paper was part of my doctoral dissertation. I thank my advisors, Leda Cosmides and John Tooby, as well as members of my dissertation committee, Daniel Conroy-Beam and Miguel Eckstein, for their advice and encouragement. For technical discussion, I thank Neil Bramley, Chentian Jiang and Chris Lucas. I am also grateful to Daniel Sznycer for sharing the behavioral data I used to calibrate the ideal observer model, and to Sean Reilly, Kathryn Marti and Francesca Tetreault for help with data collection. This work benefited from feedback by audiences at UCSB's Center for Evolutionary Psychology, ASU's evolutionary social cognition laboratory, and the annual meeting of the Human Behavior and Evolution Society.

Appendix

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

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

References

Ajzen, I., & Fishbein, M. (1975). A Bayesian analysis of attribution processes. *Psychological Bulletin*, *82*(2), 261.

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 1–10.

Barrett, H. C. (2005). Adaptations to predators and prey. *The Handbook of Evolutionary Psychology*, 200–223.

Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology*, *41*, 135–143.

Bramley, N. R., Gerstenberg, T., Tenenbaum, J. B., & Gureckis, T. M. (2018). Intuitive experimentation in the physical world. *Cognitive Psychology*, *105*, 9–38.

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708.

Cain, M. S., Vul, E., Clark, K., & Mitroff, S. R. (2012). A Bayesian optimal foraging model of human visual search. *Psychological Science*, *23*(9), 1047–1054.

Chakroff, A., Russell, P. S., Piazza, J., & Young, L. (2017). From impure to harmful: Asymmetric expectations about immoral agents. *Journal of Experimental Social Psychology*, *69*, 201–209.

Coenen, A., Nelson, J. D., & Gureckis, T. M. (2019). Asking the right questions about the psychology of human inquiry: Nine open challenges. *Psychonomic Bulletin & Review*, *26*(5), 1548–1587.

Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology*, *79*, 102–133.

Cosmides, Leda (1989). The logic of social exchange: has natural selection shaped how humans reason? studies with the wason selection task. *Cognition*, *31*(3), 187–276.

Cosmides, L., & Tooby, J. (2000). Evolutionary psychology and the emotions. *Handbook of Emotions*, *2*(2), 91–115.

Cosmides, L., & Tooby, J. (2005). Neurocognitive adaptations designed for social exchange. *The Handbook of Evolutionary Psychology*, 584–627.

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

Delton, A. W., & Robertson, T. E. (2012). The social cognition of social foraging: Partner selection by underlying valuation. *Evolution and Human Behaviour*, *33*(6), 715–725.

Delton, A. W., & Robertson, T. E. (2016). How the mind makes welfare tradeoffs: Evolution, computation, and emotion. *Current Opinion in Psychology*, *7*, 12–16.

Delton, A. W., & Sell, A. (2014). The co-evolution of concepts and motivation. *Current Directions in Psychological Science*, *23*(2), 115–120.

Dubey, R., & Griffiths, T. L. (2020). Reconciling novelty and complexity through a rational analysis of curiosity. *Psychological Review*, *127*(3), 455.

Eisenbruch, A. B., & Krasnow, M. M. (2022). Why warmth matters more than competence: A new evolutionary approach. *Perspectives on Psychological Science*, 17456916211071087.

Eisenbruch, A. B., & Roney, J. R. (2017). The skillful and the stingy: Partner choice decisions and fairness intuitions suggest human adaptation for a biological market of cooperators. *Evolutionary Psychological Science*, *3*(4), 364–378.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83.

Fisman, R., Kariv, S., & Markovits, D. (2007). Individual preferences for giving. *American Economic Review*, *97*(5), 1858–1876.

Gates, V., Callaway, F., Ho, M. K., & Griffiths, T. (2021). A rational model of people's inferences about others' preferences based on response times. *Cognition 217*.

Gelpi, Rebekah, Saxena, Nayan, Lifchits, George, Buchsbaum, Daphna, & Lucas, Chris (2021). Sampling heuristics for active function learning. *43(43)*, In *Proceedings of the annual meeting of the cognitive science society*.

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. *Oxford Handbook of Causal Reasoning*, 515–548.

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition*, *43*(2), 127–171.

Gurven, M. (2004). To give and to give not: The behavioral ecology of human food transfers. *Behavioral and Brain Sciences*, *27*(4), 543–559.

Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: Dissociable neural correlates and effects on choice. *Nature Neuroscience*, *18*(9), 1233–1235.

Hackel, L. M., Mende-Siedlecki, P., & Amodio, D. M. (2020). Reinforcement learning in social interaction: The distinguishing role of trait inference. *Journal of Experimental Social Psychology*, *88*, Article 103948.

Hamlin, K., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Developmental Science*, *16*(2), 209–226.
Heider, F. (1958). *The psychology of interpersonal relations*. Psychology Press.
Hendrickson, A. T., Navarro, D. J., & Perfors, A. (2016). Sensitivity to hypothesis size during information search. *Decision*, *3*(1), 62.
Hirschfeld, L. A., & Gelman, S. A. (Eds.), (1994). *Mapping the mind*. Cambridge: Cambridge University Press.
Hrdy, S. B. (2011). *Mothers and others*. Harvard University Press.
Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604.
Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, *123*, Article 101334.
Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. (2015). Not so innocent: Toddlers' inferences about costs and culpability. *Psychological Science*, *26*(5), 633–640.
Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, *168*, 46–64.
Jiang, C., & Lucas, C. (2021). Exploring causal overhypotheses in active learning. In *Proceedings of the annual meeting of the cognitive science society (vol. 43, no. 43)*.
Kappeler, P. M., & Van Schaik, C. P. (2006). *Cooperation in primates and humans*. Springer-Verlag Berlin Heidelberg.
Kelly, C., & Sharot, T. (2021). Individual differences in information-seeking. *Nature Communications*, *12*(1), 1–13.
Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information. *Psychological Review*, *94*, 211–228.
Kleiman-Weiner, M., Saxe, R., & Tenenbaum, J. B. (2017). Learning a commonsense moral theory. *Cognition*, *167*, 107–123.
Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2016). Looking under the hood of third-party punishment reveals design for personal benefit. *Psychological Science*, *27*(3), 405–418.
Krasnow, M. M., Howard, R. M., & Eisenbruch, A. B. (2020). The importance of being honest? Evidence that deception may not pollute social science subject pools after all. *Behavior Research Methods*, *52*(3), 1175–1188.
Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.
Lieberman, D., Tooby, J., & Cosmides, L. (2007). The architecture of human kin detection. *Nature*, *445*(7129), 727–731.
Liefgreen, Alice, Pilditch, Toby, & Lagnado, David (2020). Strategies for selecting and evaluating information. *Cognitive Psychology*, *123*, 101332.
Lim, J. (2012). *Welfare tradeoff ratios and emotions: Psychological foundations of human reciprocity*. Santa Barbara: University of California.
Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.
Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., et al. (2014). The child as econometrician: A rational model of preference understanding in children. *PLoS One*, *9*(3), Article e92160.
Lukaszewski, A. W., & Roney, J. R. (2010). Kind toward whom? Mate preferences for personality traits are target specific. *Evolution and Human Behaviour*, *31*(1), 29–38.
Markant, D., & Gureckis, T. (2014). A preference for the unpredictable over the informative during self-directed learning. In *Proceedings of the annual meeting of the cognitive science society (vol. 36, no. 36)*.
Markant, D. B., Settles, B., & Gureckis, T. M. (2016). Self-directed learning favors local, rather than global, uncertainty. *Cognitive Science*, *40*(1), 100–120.
Marr, D. (1982). *Vision: A computational approach*. MIT Press.
McKenzie, C. R. (2006). Increased sensitivity to differentially diagnostic answers using familiar materials: Implications for confirmation bias. *Memory & Cognition*, *34*(3), 577–588.
Monroe, A. (2020). Moral elevation: Indications of functional integration with welfare trade-off calibration and estimation mechanisms. *Evolution and Human Behaviour*, *41*(4), 293–302.
Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*(7031), 387–391.
Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, *118*(1), 120.
Nelson, J. D. (2005). Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*(4), 979.
Nelson, J. D., & Cottrell, G. W. (2007). A probabilistic model of eye movements in concept formation. *Neurocomputing*, *70*(13–15), 2256–2272.
Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F., & Meder, B. (2014). Children's sequential information search is sensitive to environmental probabilities. *Cognition*, *130*(1), 74–80.
Nelson, J. D., McKenzie, C. R., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological Science*, *21*(7), 960–969.
Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608.
Oaksford, M., & Chater, N. (1996). Rational explanation of the selection task. *Psychological Review*, *103*, 381–391.
Peterson, M. F., & Eckstein, M. P. (2012). Looking just below the eyes is optimal across face recognition tasks. *Proceedings of the National Academy of Sciences*, *109*(48), E3314–E3323.
Peterson, M. F., & Eckstein, M. P. (2013). Individual differences in eye movements during face identification reflect observer-specific optimal points of fixation. *Psychological Science*, *24*(7), 1216–1225.
Pietraszewski, D., & Wertz, A. E. (2011). Reverse engineering the structure of cognitive mechanisms. *Behavioral and Brain Sciences*, *34*(4), 209.
Pinker, S. (1997). *How the mind works*. Princeton University Press.
Powell, L. J. (2022). Adopted utility calculus: Origins of a concept of social affiliation. *Perspectives on Psychological Science*.
Qi, W., & Vul, E. (2020). Adaptive behavior in variable games requires theory of mind. In *Proceedings of the 42nd annual meeting of the cognitive science society*.
Qi, W., & Vul, E. (2022). The evolution of theory of mind on welfare tradeoff ratios. *Evolution and Human Behaviour*.
Quillien, T. (2020). Evolution of conditional and unconditional commitment. *Journal of Theoretical Biology*, *492*, Article 110204.
Quillien, T. (2021). *A computational framework for social valuation inference* (Unpublished dissertation), Santa Barbara: University of California.
Quillien, T., & German, T. C. (2021). A simple definition of 'intentionally'. *Cognition*, *214*, Article 104806.
Quillien, T., Tooby, J., & Cosmides, L. (in preparation). Rational inferences about social valuation in humans.
Raihani, N. J., & Barclay, P. (2016). Exploring the trade-off between quality and fairness in human partner choice. *Royal Society Open Science*, *3*(11), Article 160510.
Rothe, A., Lake, B. M., & Gureckis, T. M. (2018). Do people ask good questions? *Computational Brain & Behavior*, *1*(1), 69–89.
Ruggeri, A., & Lombrozo, T. (2015). Children adapt their questions to achieve efficient search. *Cognition*, *143*, 203–216.
Sell, A. (2005). *Regulating welfare tradeoff ratios: Three tests of an evolutionary-computational model of human anger*. Santa Barbara: University of California.
Sell, A., Sznycer, D., Al-Shawaf, L., Lim, J., Krauss, A., Feldman, A., et al. (2017). The grammar of anger: Mapping the computational architecture of a recalibrational emotion. *Cognition*, *168*, 110–128.
Sell, Aaron, Tooby, John, & Cosmides, Leda (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences*, *106*(35), 15073–15078.
Simonsohn, U. (2018). Two lines: A valid alternative to the invalid testing of U-shaped relationships with quadratic regressions. *Advances in Methods and Practices in Psychological Science*, *1*(4), 538–555.
Smith, A., Pedersen, E. J., Forster, D. E., McCullough, M. E., & Lieberman, D. (2017). Cooperation: The roles of interpersonal value and gratitude. *Evolution and Human Behaviour*, *38*(6), 695–703.
Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453–489.
Sugiyama, L. S. (2004). Illness, injury, and disability among shiwiar forager-horticulturalists: Implications of health-risk buffering for the evolution of human life history. *American Journal of Physical Anthropology*, *123*(4), 371–389.
Sznycer, D., Al-Shawaf, L., Bereby-Meyer, Y., Curry, O. S., De Smet, D., Ermer, E., et al. (2017). Cross-cultural regularities in the cognitive architecture of pride. *Proceedings of the National Academy of Sciences*, *114*(8), 1874–1879.
Sznycer, D., De Smet, D., Billingsley, J., & Lieberman, D. (2016a). Coresidence duration and cues of maternal investment regulate sibling altruism across cultures. *Journal of Personality and Social Psychology*, *111*(2), 159.
Sznycer, D., Sell, A., & Lieberman, D. (2021). Forms and functions of the social emotions. *Current Directions in Psychological Science*, *30*(4), 292–299.
Sznycer, D., Tooby, J., Cosmides, L., Porat, R., Shalvi, S., & Halperin, E. (2016b). Shame closely tracks the threat of devaluation by others, even across cultures. *Proceedings of the National Academy of Sciences*, *113*(10), 2625–2630.
Sznycer, D., Xygalatas, D., Agey, E., Alami, S., An, X. F., Ananyeva, K. I., et al. (2018). Cross-cultural invariances in the architecture of shame. *Proceedings of the National Academy of Sciences*, *115*(39), 9702–9707.
Sznycer, D., Xygalatas, D., Alami, S., An, X. F., Ananyeva, K. I., Fukushima, S., et al. (2018). Invariances in the architecture of pride across small-scale societies. *Proceedings of the National Academy of Sciences*, *115*(33), 8322–8327.
Tesser, A., Gatewood, R., & Driver, M. (1968). Some determinants of gratitude. *Journal of Personality and Social Psychology*, *9*(3), 233.
Tooby, J., & Cosmides, L. (1996). Friendship and the banker's paradox: Other pathways to the evolution of adaptations for altruism. In *Proceedings-british academy (vol. 88)* (pp. 119–144). Oxford University Press.
Tooby, J., & Cosmides, L. (2008). The evolutionary psychology of the emotions and their relationship to internal regulatory variables. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of emotions* (pp. 114–137). The Guilford Press.
Tooby, J., & Cosmides, L. (2010). Groups in mind: The coalitional roots of war and morality. *Human Morality and Sociality: Evolutionary and Comparative Perspectives*, 91–234.
Tooby, J., Cosmides, L., & Barrett, H. C. (2005). Resolving the debate on innate ideas. *The Innate Mind: Structure and Content*, 305–337.

Tooby, J., Cosmides, L., Sell, A., Lieberman, D., & Sznycer, D. (2008). Internal regulatory variables and the design of human motivation: A computational and evolutionary approach. *Handbook of Approach and Avoidance Motivation*, *15*, 251.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, *46*(1), 35–57.

Tsividis, P., Gershman, S., Tenenbaum, J., & Schulz, L. (2014). Information selection in noisy environments with large action spaces. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. (2009). Help or hinder: Bayesian models of social goal inference. *Advances in Neural Information Processing Systems*, *22*.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*, 129–140.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*, 273–281.

Wojciszke, B. (2005). Morality and competence in person-and self-perception. *European Review of Social Psychology*, *16*(1), 155–188.

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin*, *24*, 1251–1263.

Wrangham, R. W. (1999). Evolution of coalitionary killing. *American Journal of Physical Anthropology*, *110*(S29), 1–30.

Wu, C. M., Meder, B., Filimon, F., & Nelson, J. D. (2017). Asking better questions: How presentation formats influence information search. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(8), 1274.

Yu, H., Gao, X., Zhou, Y., & Zhou, X. (2018). Decomposing gratitude: Representation and integration of cognitive antecedents of gratitude in the brain. *Journal of Neuroscience*, *38*(21), 4886–4898.