University of California
Santa Barbara

A computational framework for social valuation inference

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Psychological & Brain Sciences

by

Tadeg Quillien

Committee in charge:

Professor Leda Cosmides, Chair
Professor Daniel Conroy-Beam
Professor Miguel Eckstein
Professor John Tooby

March 2021

The Dissertation of Tadeg Quillien is approved.

Professor Daniel Conroy-Beam

Professor Miguel Eckstein

Professor John Tooby

Professor Leda Cosmides, Committee Chair

March 2021

A computational framework for social valuation inference

Acknowledgements

I would like to thank my advisors Leda Cosmides and John Tooby for their support and encouragement. Their enthusiasm for foundational questions in cognitive science, and their eagerness to address these questions empirically, have been a constant source of inspiration during my transition from philosophy to experimental psychology.

While the theoretical foundations for the current work were laid out by Leda and John, the specific idea came to me while attending Miguel Eckstein's Perception course. I thank Miguel for guiding me through the first steps of creating ideal observer models. This project would not have gotten off the ground without his help. I am also grateful to Dan Conroy-Beam, who patiently endured my multiple questions about various computational techniques throughout the course of graduate school, and was kind enough to offer me space to work in his lab when I needed.

I was fortunate to collaborate with Tamsin German during my last years of graduate school. I thank her for remarking that my ideas were obvious without implying that they were trivial, which is the best form of encouragement a researcher can receive.

Many thanks also go to the undergraduate research assistants who worked tirelessly to collect the data for the studies reported here, and everyone who participated in them.

Finally, this work was made easier by the support of my friends at the CEP and the Devo area at UCSB, notably Sakura Arai, Michael Barlev, Adar Eisenbruch, Rachel Grillot, Erin Horowitz, Spencer Mermelstein, Jack Strelich, Daniel Sznycer, and Katy Walter.

Curriculum Vitæ

Tadeg Quillien

Education

2021	Ph.D. in Psychological & Brain Sciences (Expected), University of California, Santa Barbara.
2014	M.Sc. in Cognitve Science, Ecole Normale Superieure de Lyon
2012	B.A. in Philosophy, University of Nanterre

Professional Employment

Summer 2018-20	Lecturer, Department of Psychological & Brain Sciences, University of California, Santa Barbara.
2015-2021	Teaching Assistant, Department of Psychological & Brain Sciences, University of California, Santa Barbara.
2014-2015	Research Assistant, Laboratoire Dynamiques du Langage, Universite Lumiere Lyon 2 / CNRS.

Publications

Quillien, T. (2020). When do we think that X caused Y?. *Cognition, 205*, 104410. [Link]

Quillien, T. (2020). Evolution of conditional and unconditional commitment. *Journal of Theoretical Biology, 492*, 110204. [Link]

Quillien, T. (2019). Universal modesty in signal-burying games. *Proceedings of the Royal Society, B: Biological Sciences*, *286*(1906), 20190985. [Link]

Quillien, T. (2018). Psychological essentialism from first principles. *Evolution and Human Behavior*, *39*(6), 692-699. [Link]

Quillien, T. (2017). Fostering values of fairness. In T. K. Shackelford and V. A. Weekes-Shackelford (Eds.), *Encyclopedia of Evolutionary Psychological Science*. Switzerland: Springer. [Link]

Van Leeuwen, F.,**Quillien, T**., Boyer, P. (2017). Are multiple minimal outgroup males readily associated with threat? *Letters on Evolutionary Behavioral Science*, 8(1), 16-19. [Link]

Sznycer, D., Al-Shawaf, L., Bereby-Meyer, Y., Curry, O. S., De Smet, D., Ermer, E., Kim, S., Kim, S., Li, N. P., Lopez Seal, M. F., McClung, J., O, J., Ohtsubo, Y., **Quillien, T**., Schaub, M., Sell, A., van Leeuwen, F., Cosmides, L., Tooby, J. (2017). Cross-cultural regularities in the cognitive architecture of pride. *Proceedings of the National Academy of Sciences*, 114(8), 1874-1879. [Link]

Quillien, T. (2015). Population finiteness is not a concern for null hypothesis significance testing when studying human behavior: comment on Pollet (2013). *Frontiers in Neuroscience*, 9. [Link]

Awards

2017 Richard E. Mayer award for outstanding research contribution in psychology. Department of Psychological & Brain Sciences, University of California, Santa Barbara.

Research Interests

Evolutionary psychology.
Evolutionary game theory.
Computational models of cognition.
Causal cognition.

Abstract

A computational framework for social valuation inference

by

Tadeg Quillien

Organisms in a social species constantly need to make trade-offs between their own welfare and that of conspecifics. An emerging body of research suggests that the regulation of such trade-offs is an important function of social cognition. In particular, the mind has mechanisms designed to regulate tradeoffs between the welfare of the self and that of specific others, and in consequence, the mind also contains mechanisms designed to construct representations of the degree to which another individual values the welfare of the self.

Existing evidence suggests that such representations of "social valuation" play an important role in various cognitive processes such as reciprocity, partner choice, categorization and emotion. However, little is known about how people construct these representations. Because of its adaptive importance, I hypothesize that the process by which we infer social valuation is approximately consistent with normative standards of inference under uncertainty.

To test this hypothesis, I construct a Bayesian ideal observer for a simple task in which the observer, having seen the decisions made by a partner in a simple welfare-tradeoff game, needs to predict the decisions made by that partner in other rounds of the game. In a first set of studies, I find that people make predictions that closely track the predictions made by the ideal observer in that task. Additionally, participants' reports of anger toward the partner are well-predicted by the social valuation inferences made by the ideal observer, even when the different partners inflict the same opportunity

cost on the participant. I also find tentative evidence that anger ratings in that task are independently driven by deviations from expectations: individual differences in the amount by which the decisions of a partner deviated from the participant's expectations track individual differences in anger toward that partner.

In a second set of studies, I study whether people are spontaneously curious about the situations which potentially contain the most information about another person's valuation of the self. I present participants with pairs of dilemmas that another individual faced in a simple welfare-tradeoff game; for each pair, I ask them to choose the dilemma for which they would most like to see the decision that the individual had made. I find that on average, people spontaneously select the choices that have the potential to reveal the most information about the individual's valuation of the participant, in the sense of allowing the ideal observer model to draw the richest inferences.

These results strengthen the thesis that representations of social valuation are a core component of the conceptual architecture of human social cognition.

Contents

Chapter 1

Theoretical framework

1.1 Introduction

One downside of doing cognitive science is our lack of particle accelerators. We would like to break down thoughts into their elementary constituents, but have no device with which to smash them against each other at high speed until their building blocks are revealed.

Fortunately, unlike physicists, cognitive scientists study something that can be reverse-engineered. The mind is a product of natural selection, a process that engineers adaptations whose design makes sense relative to a given adaptive problem (Williams, 1966). Therefore, we can form hypotheses about the building blocks of the mind by reverse-engineering these adaptations.

In this dissertation, I investigate one candidate building block of social cognition, whose existence was conjectured from an adaptationist perspective (Tooby et al., 2008). The hypothesis is that representations of social valuation—how much agent i values the welfare of agent j—are fundamental to how we represent the (social) world. Just as we parse the world in terms of trees, rocks, buildings, and animals, we "perceive" (and care

about) the weight that individuals assign to the welfare of other members of their social world.

Before delving into the arguments in favor of this thesis, here I give a concrete example of what I mean by social valuation. This example also constitutes a simple 'demo' that allows the reader to 'see' for themselves that social valuation representation is important in social cognition. Imagine the two following situations:

a) Your housemate comes back from a long run. He is exhausted and very thirsty, but the water in your house was just cut momentarily because of maintenance work. Seeing a can of soda in your part of the fridge, he opens it and drinks it entirely, then throws it in the trash.

b) Your housemate comes back from a short walk, and steps into the kitchen. He sees a can of soda in your part of the fridge. He opens it, takes two sips, and then throws it in the trash while it is still almost entirely full.

If you are like most people, vignette (b) elicited much more annoyance than vignette (a). This feeling of annoyance came to you automatically as you read, much in the same way that some features of perception reveal themselves spontaneously in visual illusions. The difference in your emotional assessment cannot be explained by a simple tendency to negatively evaluate things that generate costs for you: Your housemate inflicted the same cost on you in both vignettes—he drank your soda. Instead, the more annoying character is the one who gained the lower benefit from his action. His action demonstrates that he is willing to inflict costs on you to gain even trivial benefits for himself (Sell, 2005). In other words, he revealed that he did not value you highly. Your emotional assessment of a situation was colored by your representation of the agent's valuation of your welfare.

In this dissertation I study the process by which you took an input (information contained in the vignette) and extracted from it an updated representation of the protagonist's valuation of your welfare. I also investigate whether it is this inference (as

opposed to other properties of the situation) that regulates your emotional assessment. Finally, I study whether people are spontaneously curious about situations with the greatest potential to reveal information about how much someone values them. Evidence that people make inferences and select evidence in a near-optimal way, and that these inferences do regulate emotional assessment, would constitute evidence for the social valuation framework.

1.2 Social valuation

What is social valuation, and why would there be cognitive adaptations to represent it? To answer this question, one must first take a close look at the logic of valuation *tout court*.

1.2.1 The concept of value

Natural selection designs nervous systems that regulate behavior in such a way that the genetic basis of these systems is likely to be replicated (Williams, 1966; Dawkins, 1976; Tooby & Cosmides, 1992). This means that one expects organisms to make decisions that are likely to lead to the kind of outcomes (e.g. acquisition of food, mating opportunities) that have a positive effect on the genetic basis of the machinery that caused this decision (Dawkins, 1982).

In a very simple organism, natural selection can achieve this aim by designing mechanisms that implement very simple decision rules such as "keep moving as long as the food concentration around you is below threshold T". However, as soon as the space of potential decisions open to an organism becomes big enough, the regulation of behavior becomes a more challenging problem, because decision-making involves trade-offs. For instance, areas richer in food may also be denser in predators, such that decisions to stay

somewhere must ideally weigh foraging opportunities against predation risk. As another example, if an organism can gain 5 units of food A by moving to the left, versus 4 units of food B by moving to the right, both courses of action preclude the gain of some amount of food (going for B means you won't get A). The organism's decision-making machinery must be able to produce adaptive decisions for classes of situations that have potential consequences for fitness.

From an idealized computational perspective, the problem of decision-making consists of generating a ranking of the different possible courses of action open to an organism in a given situation. The organism's cognitive machinery needs to be able to generate such a ranking for every situation that the organism could plausibly face. Natural selection's problem is to design an organism whose nervous system implements an approximation of such a ranking system (where possible actions are ranked according to their expected consequences for the replication of organismic design[1]).

Of course, for any but the simplest organisms, it is impossible to create a nervous system which explicitly represents such a ranking (by enumerating every possible {situation, possible actions} set and assigning a rank to each action for each set). Instead, organisms need to be able to compute 'online' which action is best for the situation they are currently facing.

In this respect, the problem of decision-making is analogous to the problem of linguistic competence. There is an infinity of possible things you could say, and therefore it is impossible to explicitly encode every single sentence you could utter in a big list stored somewhere in your brain. Language is able to generate a potential infinity of sentences from a finite cognitive architecture because of its compositional nature (Pinker, 1994): our mind contains rules for combining building blocks (e.g. words) into bigger units (e.g. sentences).

[1]In the statistical sense of 'expected'. See also footnote 3.

There is an infinity of possible decision problems a complex organism could be faced with. This suggests that decision-making must rely on a combinatorial system, which combines basic features of a situation, in order to generate a ranking of possible actions on the fly.

An agent can implicitly represent a ranking over possible actions by using a function which maps every possible action to a numerical value.

For instance, let us consider an organism with a simple life cycle. The organism is born, and all it has to do in its life is make one decision:

-option 1: get W units of food A and X units of food B

-option 2: get Y units of food A and Z units of food B

where W, X, Y, Z are continuous variables whose values vary from one individual to the next. Then the organism reproduces, and its expected number of offspring is proportional to $\sqrt{A}+\sqrt{2B}$ (i.e. while food B tends to have a larger effect on fitness than food A, the organism still benefits from a balanced diet).

The number of possible decisions that the organism could face is infinite (all possible combinations of W, X, Y, Z), so natural selection cannot explicitly hard-wire a recommended action for each possible dilemma. Instead it is likely that the species will evolve the following decision rule:

-For each option, compute the 'value' of that option according to the formula $V = \sqrt{A} + \sqrt{2B}$

-choose the option with the highest value.

This is in fact the optimal decision rule with respect to fitness maximization[2]. It takes as building blocks various features of the expected outcomes of the decision (here:

[2]In the simple example I use here, there is an infinity of other decision rules that would do equally well: for example $V = \sqrt{A} + \sqrt{2B} + 42$. However, if the organism had to make probabilistic decisions (decisions involving options like "get X units of food A with probability p, and nothing otherwise"), then the only optimal decision rules would be linear transformations of $V = \sqrt{A} + \sqrt{2B}$.

amount of food A, amount of food B) and combines them to generate a 'value' representation, which is then used to determine choice. This way of generating decisions has various desirable properties, for instance, it generates decisions that are in a certain sense consistent with each other (see Von Neumann & Morgenstern, 1953; Savage, 1954).

The upshot of this argument is that an ideal decision-maker will make decisions by (i) computing a representation of the possible consequences of the possible actions it could take (what economists call the "feasible set"), (ii) on the basis of these representations, computing the value of each possible action, (iii) implementing the action that has the highest value (Von Neumann & Morgenstern, 1953; Savage, 1954). The concept of 'value' can therefore be seen as emerging as the solution to the information-processing problem of ranking different possible outcomes when making decisions.

Here it is useful to remind the reader that the current discussion is taking place at the computational level of analysis. I am not claiming that actual organisms necessarily engage in explicit value computations, or that their decisions will implement a perfectly consistent implicit ranking of possible actions (for an argument against the idea that the human brain explicitly engages in value computations, see Hayden & Niv, 2020). The arguments above merely mean that, to the extent that natural selection crafts well-designed decision mechanisms, one generally expects organisms to behave *as if* they were engaging in value computations. That is, it should be possible to approximately predict an organism's behavior by assuming that its nervous system is explicitly assigning numerical values to possible outcomes (see Dawkins, 1976; Parker & Maynard-Smith, 1990 for the heuristic value of modelling organisms as explicitly solving optimization problems; and Friedman, 1953 for a similar argument in economics).

In sum, it is helpful to think of organisms as assigning a value to certain outcomes: they will assign high value to acquiring the food they need, getting access to mates, prevailing over a rival, and so on, and negative value to not having water, getting close

to predators, etc [3]. As a corollary, it is helpful to think of organisms as representing possible actions in terms of their 'costs' and 'benefits'. Doing so is not only helpful to scientists who study animal behavior; organisms themselves can predict the behavior of other organisms by representing them as assigning values to outcomes[4].

1.2.2 Selection pressures for welfare-tradeoffs

The fact that most organisms share their world with other organisms greatly complicates decision-making. To a large extent, each organism will have its own idiosyncratic valuation system, since the genetic interests of different individuals[5] are almost always non-identical (Williams, 1966; Trivers, 1974).

As a consequence, a variety of selection pressures lead one to expect that decision-making in many organisms will rely on a more complicated system of valuation than that suggested in the previous section. This system of valuation would often need to make some value computations from the point of view of another organism.

More specifically, a valuation system designed for life in a complex social world would likely rely on different types of reference frames for valuation. The simplest reference frame is a self-centered reference frame: it is the one according to which more food, safety, health, etc, to the self are good things.

But there would also be an altercentric reference frame: it is the one according to which food, safety, etc, to someone else (e.g. a sibling) are good things. Finally, one also expects the existence of a meta-representational reference frame, which is a representation of the self-referential valuation system of someone else.

[3]Note that these value assignments will not always be 'correct' in the sense of assigning the highest value to the outcomes that most promote the organism's inclusive fitness. Rather, natural selection designs valuation systems that tend to promote fitness on average, across all individuals equipped with this system, in the environment in which the system has evolved (Tooby & Cosmides, 1990).

[4]In humans, 'cost' and 'benefit' seem to be early-developing conceptual primitives of commonsense psychology (Liu, Ullman, Tenenbaum, & Spelke, 2017).

[5]Or, for that matter, genes within an individual (Cosmides & Tooby, 1981; Dawkins, 1982).

Kin selection is an important selection pressure for an altercentric reference frame for valuation. A gene can promote the replication of the design it codes for by acting on the replication of copies of that gene in close relatives of its bearer. As a result, natural selection tends to design individuals that care for the welfare of their close genetic relatives (Hamilton, 1964)[6] .

In some organisms, simple decision rules may be enough to ensure that they behave in a way that weighs the welfare of their relatives appropriately. However, in many species, selection has to design mechanisms enabling an individual to recognize kin, and represent potential actions in terms of their payoffs for the relevant kin members (Lieberman, Tooby & Cosmides, 2007). Then the individual's decisions will be a function of payoffs for oneself as well as payoffs to the relatives affected by one's actions. In an idealized case where genetic relatedness is the only factor influencing welfare-tradeoffs, and under certain other background assumptions (McElreath & Boyd, 2007), the individual will compute the value of a state of the world as:

$$V = v_{self} + r_{est} * v_{other}$$

Where v_{self} is the action's value for the self, v_{other} the action's value for its relative, and r_{est} is the kinship index (Lieberman, Tooby & Cosmides, 2007) that the individual assigns to its relative. Here, the function that computes vother has been optimized by natural selection to assign a high value to states of the world that were likely to maximize the likelihood of replication of the relative's genes (in the EEA). v_{other} belongs to the individual's altercentric reference frame for its relative. Note that it does not depend at all on the relative's own system of valuation. In other words, kin selection does not necessarily require that an individual meta-represents the value systems of others. For

[6]See link and link for tutorial agent-based models that illustrate the logic of kin selection.

instance, an adult who denies his daughter extra ice cream out of concern for her health is motivated by his perception of the actual costs to his child (his altercentric reference frame for the child). If he were sensitive to the child's valuation system, he would give her as much ice cream as she wants.

In contrast to kin selection, other selection pressures for welfare-tradeoffs often require one to infer what another person will value, and store this information in a metarepresentation. These selection pressures have to do with the fact that the behavior of others might be influenced by your actions, and that their behavior will depend on the costs and benefits they perceive.

Reciprocity (Trivers, 1971) is a paradigmatic example. Suppose that an individual's social behavior is regulated by a strategy of the type "if you deliver a benefit to me, I increase my propensity to deliver benefits to you". This strategy can be evolutionarily stable, provided that the probability of future interactions is large enough, and that delivering benefits to others is not too costly (Axelrod & Hamilton, 1981), because individuals with the strategy disproportionately help other individuals with the strategy. Mathematical studies of reciprocity often assume for convenience that the fitness costs and benefits of actions are transparent to every agent, but in reality this condition is unlikely to hold most of the time. One expects that reciprocity would act as a selection pressure for organisms to meta-represent the valuation systems of others, so that they are able to take actions that get registered as genuine instances of benefit delivery by the recipient.

Note that I am not suggesting that the ability to meta-represent the valuation system of others is a *necessary* requirement for the evolution of reciprocity, or any other form of cooperation. Indeed, reciprocity can be sustained via simple decision rules in organisms without a complex nervous system (see e.g. tit-for-tat interaction between fungi and plants, Kiers et al., 2011). Rather, the claim is that for organisms (such as humans)

with a complex social life (e.g. organisms who can benefit from exchanging an open array of goods and services), selection for reciprocity should also favor the evolution of sophisticated mechanisms for inferring and meta-representing what others value. As Cosmides (1985, chapter 5) succinctly puts it, "to engage in an exchange with you, I must know what you want".

Reciprocity can also take the form of 'negative' reciprocity, a.k.a. punishment (Clutton-Brock & Parker, 1995; Morris, MacGlashan, Littman & Cushman, 2017): if you inflict costs on me, then I will inflict costs on you. In a social ecology where retaliatory punishment is part of the behavioral repertoire, individuals have an incentive to minimize the costs they inflict to others, all else being equal.

In humans, partner choice is also an important selection pressure for welfare-tradeoff mechanisms (Baumard, Andre & Sperber, 2013; Barclay, 2013). Individuals tend to seek out interactions with people who deliver the best payoffs. As a way to ensure access to such people, a useful strategy is to make sure that valuable partners associate oneself with positive payoffs.

Somewhat paradoxically, antagonistic interactions (i.e. conflict over a resource) also may require welfare-tradeoffs. Consider two birds that both want the same piece of food. Since the birds are competing, there is a sense in which each bird negatively values the welfare of the other (e.g. it would be better for bird A if bird B suddenly had a heart attack). However, it follows from the logic of animal conflict that one expects each bird to behave as if it assigned positive value to the other's welfare. Evolutionary game-theoretic models of conflicts (Maynard-Smith & Parker, 1976; Hammerstein & Parker, 1982; see also Sell, 2005) predict that in many contexts, the optimal strategy is to adjust one's likelihood of engaging (and continuing) in a fight as a function of how much one values the resource relative to how much the other contestant values it. In other words, individuals are often better off yielding a resource if it is worth much more to the contestant than it

is to them. Extortion ("if you do not give me this benefit, I will inflict costs on you") is another example where it is sometimes optimal to behave as if one valued the welfare of one's antagonist. When a robber says, "your money or your life!", it is prudent to make a decision that puts weight on the robber's welfare. A transfer of money from oneself to the robber should be seen as highly valuable when the alternative is a bullet to the head. These examples (conflict over resource, and extortion) are situations where the organism may assign a negative value to the other organism's welfare in its altercentric reference frame, but still meta-represent that organisms' valuation system and make choices that are constrained by it[7].

1.2.3 The form of welfare-tradeoff mecahnisms

Given the selection pressures reviewed above, one can make conjectures about the architecture of the cognitive mechanisms that determine welfare trade-offs.

The existence of different reference frames for valuation makes it unlikely that Alice's mind would always use a catch-all category for "payoffs to Bob". It is more plausible that (at least in some cases), she would keep her altercentric and meta-representational reference frames for Bob distinct. For instance, imagine that Alice's brother, Bob is a drug addict. In her altercentric reference frame, 'Bob gets 10g of heroin' is a cost, since she cares about Bob's health; but in her meta-representational reference frame it is a benefit (it is likely that Bob wants to get 10g of heroin). In order to behave optimally, Alice needs to integrate these two values in the correct way. Subsuming these two values within a single reference frame is suboptimal: for instance if she computes the value of the event as the sum of its altercentric and metarepresentational values, these values

[7]Even experts in military strategy are sometimes surprised at the simple insight that enemies usually share common interests, such as the avoidance of complete mutual destruction (as observed by Thomas Schelling in his preface to the 2nd edition of The Strategy of Conflict, 1960/1980). The tension between altercentric and meta-representational reference frames for valuation may explain why this insight is somewhat counter-intuitive.

would cancel each other and she would remain indifferent to the event. Instead she needs to prevent the event, while accounting for the fact that this will be met by resistance by Bob.

Conceptually, it is helpful to think in idealized terms, and conceive of welfare tradeoffs as being regulated by a valuation function. This valuation function assigns an overall value to a given state of the world by integrating the value of that state of the world from the point of view of various agents (using different reference frames). It is likely that the general form of this valuation function is a reliably-developing part of the human cognitive architecture, and is shared by most people. However, the parameters that regulate specific settings are expected to vary across the individuals making the tradeoffs, and across the individuals that are the targets of this tradeoff; for instance, some people are nicer than others, and Alice loves her mother more than she does a random stranger. Here I will call these parameters Welfare-Tradeoff Parameters (WTPs).

It is likely that there are many such WTPs. For instance, a basic prediction of reciprocity theory is that people will assign a higher value to the welfare of others when their decisions are observed: this implies the existence of at least one Welfare-Tradeoff Parameter that determines how much more generous an individual is when observed compared to unobserved. A basic prediction of models of animal conflict is that an individual's formidability (i.e. ability to inflict costs) will often influence how much you value their welfare, meaning that there should also be a set of parameters that regulate how much your decisions change as a function of a person's formidability. By continuing to list all relevant selection pressures and taking into account the interactions between them, we would end up writing a valuation function with very many parameters. Such a task would be too difficult, so instead we will make the problem tractable by deliberately pretending that it is simple.

1.2.4 Welfare-Tradeoff Ratio: a toy model of welfare-tradeoff psychology

We do not live in a frictionless world of point particles, yet physicists often pretend otherwise for their calculations. Here we will follow their lead and deliberately consider a simplified model of human welfare tradeoff psychology. Specifically, assume that Alice's valuation function[8] is:

$$V = V_{alice} + WTR * V_{bob}$$

Where WTR is Alice's Welfare Tradeoff Ratio toward Bob: it is the exchange rate at which she trades off Bob's welfare against hers (Sell, 2005; Delton, 2010).

There are two benefits to such a toy model. One is conceptual, the other is methodological.

On the conceptual side, a simplified computational model strikes a necessary balance between two extremes. At one extreme, we often use folk psychology to explain people's behavior. Folk psychology is (almost by definition) intuitive: when fed information about an agent, it automatically generates a deluge of inferences. Its downside is that these inferences do not come with deep causal justifications. On the other extreme, complex computational models can generate rich and flexible predictions across a variety of cases. Their downside is that they are extremely unintuitive, even to the people trained in them: short of plugging in numbers into the formula, they cannot be used to gain a deep intuitive understanding of the relevant computational principles.

The WTR formula strikes a middle ground. It sweeps under the rug many complications, yet it retains the rigor of an explicitly computational model. Because of its

[8]For added simplicity, we also ignore the question of whether V_{bob} refers to Alice's altercentric or metarepresentational reference frame for Bob.

simplicity, it is easy to see what each component does, and to derive predictions. One can see the WTR model as being essentially an operation of dimensionality reduction: take the highly multidimensional architecture of social valuation, and reduce it to a single continuous variable, with a straightforward interpretation: an exchange rate between your welfare and mine.

The WTR valuation function is also a useful tool for modelling human decisions in simple situations. For example, in a task where the participant's decisions are always unobserved, one can model people's behavior with a model that does not need to include parameters for the effect of observability on behavior. By designing a sufficiently simple task, we can create a simple mathematical model that has a chance of successfully accounting for human behavior in that task. For instance, Delton (2010) introduced the Welfare Trade-off Task (WTT). 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 0

Or

Bob receives $\$\pi_{bob}$ and Alice receives 0

The dictator plays several trials of the game. Across trials, the value of π_{alice} varies, while π_{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.

Empirically, because the task is so simple, it is possible to model human inferences about valuation with the WTR model described above. Another advantage of such simple experiments is that they can be used to evaluate basic characteristics of the welfare-

tradeoff machinery. For instance, well-designed decision-making mechanisms should generate decisions that are internally consistent (for example, if one of your decisions reveal that you (strictly) prefer apples to oranges, and another decision in the same context reveals that you prefer oranges to apples, your decisions are suboptimal). Delton (2010) showed that people's decisions in the WTT were highly consistent in that sense, at least when the amount the other will gain is relatively constant (varying within a small range), but the opportunity cost of giving that amount varies. Other researchers found that the choices of individuals are internally consistent even when the price of giving versus keeping money varies continuously, and the amount to divide changes (e.g., Andreoni & Miller, 2002; Fisman, Kariv & Markovits, 2007).

A terminological note. In this work, I will try to use the term *WTPs* when talking about social valuation in general. I use the term *WTR* within the context of experimental tasks (such as the WTT) in which a one-parameter valuation function is a good model of human behavior. When I discuss the result of these experiments, to maintain a sense of consistency with the Methods and Results section I will say things like "people can apparently infer the WTR of others"; this should be interpreted as saying that people infer those WTPs (whatever form they have) that are relevant to predicting behavior in the WTT. I may also use *WTR* in its role of a simplified conceptual model when it is convenient to conceive of social valuation as unidimensional (e.g. "anger is triggered by cues of low WTR").

Some of the selection pressures for welfare-tradeoffs listed above highlight an interesting fact. In many circumstances, it may be adaptive for Alice to adjust her welfare-tradeoff parameters (WTPs) toward Bob as a function of Bob's WTPs towards her. It may also be adaptive for Alice to adjust her WTPs as a means to influence Bob's WTPs towards her. This raises the possibility that people have adaptations to infer the magnitude of other people's WTPs.

1.3 Social valuation inference

1.3.1 Adaptive problems whose solution require social valuation inference

The most straightforward use of social valuation inference is prediction. In general, knowing what other individuals value is a good guide to their future actions. Knowing how much they value the welfare of others is useful for predicting how they will behave towards them. For instance, Alice can predict how much of his cake Bob will share with her if she has an accurate assessment of Bob's WTPs towards her.

Predicting someone's welfare tradeoffs is especially important in order to estimate the payoffs (i.e. costs and benefits) they will deliver to you. Estimating the expected payoffs that Bob will deliver to her helps Alice to decide whether to seek Bob's company or avoid him, and whether Bob's existence is instrumentally beneficial to her.

Of course, social valuation inference is not strictly speaking necessary for solving this estimation problem. Instead, Alice may predict the payoffs that Bob will deliver to her in the future by simply computing the average payoffs that Bob has delivered to her in the past, and using this value as her best estimate. Following Lim (2012), we will call this model the "net profit model". The net profit model is far from optimal. For instance, suppose that so far, Bob always had the opportunity to help Alice at low costs to himself; if his situation changes such that it becomes costlier to help Alice, it is likely that he will deliver fewer benefits to Alice than he did so far. The net profit heuristic is incapable of making this prediction.

Instead, Alice's best bet is to build a causal model of the factors that influence Bob's payoff delivery to her (Barrett, Cosmides & Tooby, 2010). This model must include a mix of 'external' factors (e.g. how costly is it on average for Bob to help Alice) as well as facts

about Bob's psychology (how much control he has over his actions, his knowledge of the consequences of his actions on Alice's welfare, etc); as part of these psychological facts, Bob's WTPs towards Alice are a crucial component. Given the appropriate background knowledge, a well-calibrated causal model allows much more flexible predictions, with the need for much less data, than simple extrapolation (Pearl, 2000).

A second use of social valuation inference is intervention. Having an accurate representation of a variable in the world makes us more effective at intervening on the value of that variable. For instance, to ensure that our body has the right amount of water, we have an internal estimate of water need (which we subjectively experience as thirst) that allows us to regulate how much immediate effort we put into acquiring water (see Tooby et al., 2008). Similarly, we are more effective at setting the WTPs of others at the 'correct' values (from our point of view) if we can accurately estimate them.

In particular, it is likely that natural selection designed adaptations that enable us to compute, for a given situation, the level at which we 'deserve' to be treated. For instance, in a biological market, supply-and-demand forces determine the share that an individual can expect to receive from the fruits of a joint collaboration, given the individual's productivity, the market's fluidity, etc (Debove, Andre & Baumard, 2017). One expects an individual to have an internal representation of the share they can reasonably expect to be offered, given (e.g.) their productivity. If they get an offer that is lower than this value, it is a good bet that asking for a better offer will work. To take another example, in reciprocal exchange, individuals adjust their level of cooperation to that of their interaction partner. A strategy that represents a partner's 'level of cooperation' as simply whether that partner recently cooperated or defected may fare poorly, because it would construe as "defection" instances where one's partner fails to help because it was impossible or too costly to do so. Instead, a more plausible cognitive architecture for reciprocity is one where interactants adjust their WTPs as a function of their estimate

of the WTPs of their partner (Lim, 2012). Given the WTPs that you express toward your partner, you expect them to have certain WTPs towards you. If they do not, then lowering your WTPs (or threatening to) may be an effective way to recalibrate your partners'.

There are many ways we can attempt to recalibrate someone else's WTPs. We can punish and reward, issue threats and promises. We can communicate traits or intentions that reliably advertise us as someone whose welfare should be valued. These may be our formidability, productivity, possession of certain skills, shared interests with the partner, ability to generate positive externalities, or commitment to that relationship (Sell, Tooby & Cosmides, 2014; Sznycer et al., 2017; Tooby & Cosmides, 1996; Quillien, 2020a). Regardless of which of these strategies we use, we will generally have a better chance of efficiently recalibrating the target's WTPs when we estimate them accurately. Overestimating the weight that someone else puts on your welfare may lead you to neglect opportunities to recalibrate their WTPs to your advantage. Conversely, underestimating that weight may lead you to try to recalibrate the WTPs of others when the attempt is unlikely to succeed. Indeed, excessive eagerness to bargain for better valuation may endanger existing relationships.

Machinery for social valuation inference may also serve a meta-meta-representational function. Being able to predict the kinds of social valuation inferences that others will draw may often be useful.

For one, this ability is useful when we explain to others why we drew the inferences we did. When we are angry at someone, we might explain why by emphasizing the magnitude of the cost that the target of our anger inflicted on us, for example. In order for such explanations to be successful, we need to be able to recognize whether they will sound acceptable to the target: upon hearing our explanation, will they conclude that our anger makes sense? The goal of an explanation for anger is to explain to the target

why we think they do not value our welfare enough. In order to do that, we must be able to correctly identify the elements of the situation that were causally relevant to our anger, that is, the elements that caused us to infer low valuation (for instance, the magnitude of the cost we incurred, but not the color of the transgressor's t-shirt). If they conclude that the explanation for our anger is well-formed, the target can recalibrate her WTPs accordingly, or provide information ("I did not know this would hurt you that much") that clarifies their behavior.

Second, to a large extent our own decisions should be a function of the inferences people make about our WTPs. All else being equal, we should do more of the things that make us look good. Doing so requires a good model of the inferences people make. Such a model also helps us convince people that we value them highly, by highlighting relevant episodes from the past that would be highly effective examples (e.g. "of course I care about you, remember when I woke up at 4am to give you a ride to the airport!").

Third, it will sometimes be useful to strategically exploit representations of the WTPs of a third-party. Pointing out that the village chief recently gave us a huge favor may help our claim for status. Exaggerating the offense that a member of the rival village committed toward one of us may help coordinate our village towards a raid. In order to do so effectively, we need a good model of how the minds of others will draw WTP inferences from the information we give them.

So far in this section, I have been arguing that inferences about the WTPs of others are potentially useful for a variety of adaptive problems. Recently, Eisenbruch & Krasnow (2019) made the stronger argument that inferences about WTPs are *more useful* than inferences about other traits (such as competence-related traits). Their argument relies on extensive empirical evidence that the statistical distributions of WTPs and competence traits differ in important ways. 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, 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 if 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 provide support for the logical validity of their argument with evolutionary agent-based models simulating a partner choice process. They find that natural selection designs agents that preferentially attend to another agent's generosity rather than to its productivity when the former has higher between-agent variance but lower within-agent variance.

1.3.2 Social valuation inference and social emotions

As seen above, inferences about social valuation are likely to have a wide range of functional consequences. In order to coordinate the wide range of functional responses appropriate for a given situation, natural selection has designed modes of operation that are commonly referred to as emotions (Cosmides & Tooby, 2000).

For instance, inferring that someone values you less than you expect should lead to efforts to recalibrate that person's WTPs, which requires a host of coordinated responses:

signal that you inferred a low valuation, explain why you made this inference, explain why the valuation you inferred is inappropriate, signal that you may take action (inflict costs, withhold benefits, etc) contingent on successful recalibration on the part of the target, credibly signal your ability and intention to take such actions, ready your body to take these actions in case they become necessary, etc. By hypothesis, anger is the emotion that coordinates these functional responses (Sell, 2005).

An inventory of the design features of human anger is strongly suggestive of the emotion's recalibrational function (see in-depth review in Sell, 2005). Furthermore, the theory has made successful novel empirical predictions. Physical strength is a cue of one's ability to inflict costs in response to low valuation. Accordingly, the typical features of the human anger face increase the perceived physical strength of the angry person (Sell, Tooby & Cosmides, 2014). Physically strong males are also more prone to anger (Sell, Tooby & Cosmides, 2009). Evidence also suggests that social valuation inference is an input to the emotion (see next section).

Gratitude may be an emotion designed to adjust one's WTPs, and communicate relevant information to one's benefactor, after receiving a benefit from someone else. It is likely that one of the inputs to gratitude is the magnitude of the benefit delivered. The recipient should communicate the magnitude of the perceived benefit, as feedback to the benefactor, in order to guide him toward similar beneficial actions in the future.

Another input may be the inference that the benefactor has a high valuation of the recipient's welfare (Sznycer, 2010; Lim, 2012). The latter hypothesis stems from the logic of mechanisms such as partner choice and reciprocity. For instance, assuming that human reciprocity is based on updates of one's WTPs as a function of the WTPs of one's partner (Lim, 2012), then gratitude may be the emotion that mediates such updating. Also, in order to keep benefits from Alice flowing towards him, Bob needs to convince her that her beneficial acts are causally effective in maintaining / increasing his valuation of

her. The best thing he can do to convince her of this is to deliver benefits to Alice, but short of immediate opportunities of doing that, he can send signals that he acknowledges that Alice puts a high weight on his welfare[9].

A potentially important aspect of gratitude is also that it creates common knowledge between Alice and Bob that Alice values Bob. Indeed, to the extent that there is no ambiguity about whether Alice knows that Bob knows that Alice values him (and so on), their reciprocal relationship is on more solid grounds, which is beneficial to both.

Empirical data confirm that gratitude mediates the upregulation of our WTPs after receiving a benefit that is diagnostic of someone's valuation of our welfare (Lim, 2012; Smith et al., 2017). Expressions of gratitude are also interpreted by the benefactor as indicating a higher likelihood of future altruism on the part of the recipient, and even children make this inference (Thomsen et al., 2018).

Other social emotions may be designed to take inferences about social valuation as input. For instance, guilt may be an internal signal that one has placed too low a weight on someone else's welfare (Tooby & Cosmides, 1990; Sznycer, 2019). Shame and pride are internal signals that one's traits or deeds have negative (for shame) or positive (for pride) consequences for how people value us (Sznycer, 2019)[10].

The hypothesis that social emotions take as input the output of social valuation inference mechanisms suggests that self-report of felt emotion can be an indirect way

[9]Some WTP theorists (e.g. Sznycer, 2010; Lim, 2012) consider that gratitude is activated by acts that reveal that Alice has a higher valuation of Bob than Bob expected. This may not necessarily be the case. For instance, people still express gratitude when receiving presents from people they have known their whole life. Expressing gratitude in response to a benefit delivery is potentially useful even when it acknowledges a valuation that is unsurprising to you. This is because failure to acknowledge it may lead your partner to doubt that their helpful actions are appropriately recognized, and subsequently decrease the weight they put on your welfare.

[10]Note that in the current work I use 'social valuation' in the narrow sense of one's propensity to make welfare trade-offs. The proper domain of shame and pride is probably social valuation in a broader sense: for instance one may be proud of one's physical attractiveness because it increases one's mate value, independently of (and in addition to) any potential effects it has on how people weigh our welfare. But since welfare valuation is an important subset of social valuation writ large, it is highly relevant to shame and pride.

of studying these inferences. If emotional reports are regulated by the same cues that should elicit inferences, this is evidence that they take the result of such inferences as input (and by the same token, that people do make such inferences). In chapter 2, I will use this strategy with respect to anger and gratitude. The theoretical arguments reviewed so far lead us to expect that people make inferences about the WTPs of others. Do they?

1.3.3 Evidence that people make social valuation inferences

Existing evidence comes from studies that show that the net profit model is insufficient to account for many aspects of human psychology. Instead, human behavior is often sensitive to what can be interpreted as cues of social valuation.

Many of these studies employ a similar logic: participants play an economic game with partners that differ along two dimensions, which are orthogonally manipulated by the experimenter. The first dimension (which we will call 'productivity') is how much reward the partner has the potential to deliver to the participant. The second dimension (which here we will call 'WTR') is the tradeoff the partner makes between their welfare and that of the participant. The typical result is that the latter dimension matters much more than the first in how participants perceive their partner, whether they want to interact with them again, and their emotional reactions toward them. This is the case even in situations where a rational payoff-maximizing agent would weigh productivity more highly.

Lim (2012) showed participants two partners who had completed the WTT in the role of the dictator, with the participant as recipient. The choices made by the two (sham) partners, as well as the payoffs in the WTT that they played, were manipulated such that the first partner expressed a WTR of .9 to the participant while the second

partner expressed a WTR of 0. However, across all trials of the WTT the expected value of the benefit to the participant from the second partnerwas twice as large as the expected value from the first partner. 83% of participants preferred to interact with the high-WTR, low-profit partner in the next round. Participants also expressed a larger WTR to the high-WTR partner when they subsequently played in the dictator role, and they expressed more gratitude and less anger toward the high-WTR partner. They also mistakenly perceived the high-WTR partner as having delivered a higher reward to them than the low-WTR partner[11]. By contrast, in other conditions where the two partners differed in productivity but expressed the same WTR as each other, productivity had very weak effects on these dependent measures.

Hackel, Doll & Amodio (2015) found similar effects using a slightly different design. They had participants interact with four different (sham) partners in a series of simple dictator games. Partners differed in their endowment, and in the fraction of their endowment they decided to share with the participant, with the two factors independently manipulated. In each trial, the participant could choose to play (in the receiver role) with one among two of the partners; after choosing a partner they could see both that partner's endowment and the percentage they shared. Over the course of many interactions, the participant could learn the WTR of each partner, as well as the average rewards that partner delivers. In this task, the reward-maximizing strategy is to ignore your beliefs about WTR and simply choose the partner who delivered more rewards in the past. However, Hackel et al. found that people placed a greater weight on WTR rather

[11]Although exploratory, this result is interesting, since 'mistakes' of this kind often happen at lower levels of perception, for example in vision. For instance, when looking at the Adelson checkerboard, people mistakenly think that square B is lighter than square A, even though both squares have the same luminance. They make that mistake because their brain rationally infers that the reflectance of square A is darker. In other words, because luminance is a cue to reflectance, and the perceptual system is designed to infer reflectance, people confuse the cue for what it is a cue of. It is possible that the same thing is happening in Lim's experiment: by hypothesis, the brain is designed to infer WTR, and under normal conditions the amount of profit one gets from someone is a cue to their WTR. As a consequence, when asked about profit, people answer by giving their WTR estimate instead.

than expected reward when choosing partners. When asked who they would most like to play with in a subsequent, unrelated, cooperative puzzle-solving task, they also based their choices mostly on the partners' WTR. Information about WTR was encoded in the ventral striatum (a brain region associated with reward processing), as well as brain areas (such as the right temporoparietal junction) involved in impression formation. Integration of WTR information for partner choice decisions involved the ventro-medial Prefrontal Cortex (vmPFC).

In a follow-up experiment using a similar design, Hackel, Mendle-Siedlecki & Amodio (2020) found that people relied on WTR information more when choosing among people than choosing among slot machines with the same objective characteristics (endowment size and sharing 'behavior'). The fact that cues of agency increases the importance of 'WTR information suggests that latter matters because it is treated as a cue of social valuation.

The primacy of WTR information for partner choice and impression formation was also conceptually replicated by Raihani & Barclay (2016) and Eisenbruch & Roney (2017). In Raihani & Barclay (2016), participants witnessed the decisions that two dictators had made in a previous game. The dictators varied in the size of their endowment ($.5 vs $2.5) and the share they had given to the former recipient ($20% vs 50%). Participants had to choose which of the two players they would like to play a game with next (in the receiver role). They were told that there was a 90% chance that the dictator's endowment would remain the same in the next round. Given this, and assuming that a dictator's WTR remains constant across rounds, the expected payoff of choosing a rich but stingy partner is $.2 * (.9 * 2.5 + .1 * .5) = \$.46$, while the expected payoff of choosing a poor but fair partner is $.5 * (.9 * .5 + .1 * 2.5) = \$.35$. Therefore, when faced with a choice between a poor-but-fair and a rich-but-stingy partner, a payoff-maximizing agent will choose the

rich-but-stingy partner[12]. Nonetheless, 57% of choosers chose the poor-but-fair partner (although this was not significantly different from 50%).

Eisenbruch & Roney (2017), employed a design with a similar logic, except with a trust game. Participants were given a $10 budget, and chose to send any amount they would like to a partner. Any amount that the partner would return would be multiplied by 3, 4 or 5. This multiplier, which varied across conditions, was the 'productivity' of the partner. Partners were sham players, who returned either 30%, 40%, or 50% of the amount they were trusted with. Then, participants indicated whether they would like to play further rounds of the game with the partner. From an economic point of view, the productivity and the WTR of a partner make equal contributions to the rewards generated by that partner, so the net profit model predicts that neither should be given preferential treatment. Yet the effect of WTR on partner choice decisions was more than 4 times larger than the effect of productivity.

Other evidence for social valuation inference comes from studies on emotion. According to some theories, emotions such as anger and gratitude function to regulate the WTPs of others. For instance, anger communicates that the target does not value us highly enough (Sell, 2005), while gratitude communicates that we acknowledge that an act by the target reveals that they value us highly (Lim, 2012; Smith et al., 2017).

Across several studies (including the one reviewed above), Lim (2012) consistently found that agents whose decisions in a WTT expressed a high WTR elicited more gratitude and less anger than agents whose decisions expressed a low WTR, even when holding benefit delivery constant, or when the low-WTR agent delivered more benefits.

Aaron Sell and his colleagues have conducted a wide range of studies testing various predictions of the recalibrational theory of anger (e.g. Sell, Tooby & Cosmides, 2009;

[12]Note that if we relax the assumption that the dictator's WTR level remains the same across rounds, it is even more rational, in economic terms, to choose the rich-but-stingy partner.

2014). The studies that most directly speak to the topic of social valuation inference are reported in Sell (2005) and Sell et al. (2017). The authors asked participants to read vignettes in which a perpetrator inflicted a cost on the participant in order to receive a benefit (for instance, cut in line at a public telephone booth in order to prevent a winning lottery ticket from being lost). In several studies, they manipulated the benefit to the perpetrator, the cost to the participant, as well the perpetrator's intention (whether they intended to inflict the cost on the participant in particular). All these variables had the predicted effect: anger was most strongly elicited by perpetrators who inflicted a large cost on the participant, did so intentionally, and gained a low benefit from their action. These results held across different cultures including participants from a small-scale society in the Ecuadorian amazon, the Shuar (Sell et al., 2017).

It has been known for a long time that actions that are more costly to the benefactor, more beneficial to the recipient, and where the benefactor intends to benefit the recipient, elicit more gratitude (Tesser et al., 1968; see also Yu et al., 2018). Cost and intention are important cues to social valuation[13]: a benefactor who knowingly pays a large cost to benefit the recipient reveals that he sees the benefit to the recipient as valuable enough to compensate a large cost to himself, and the fact that he did so intentionally[14] warrants the validity of the inference. These data, as well as the effects of expressed WTR on gratitude found by Lim (2012) support the idea that inferences about social valuation

[13]The extent to which the magnitude of benefit delivery is a cue to social valuation is more ambiguous. On the one hand, large benefits to the recipient (with cost to the benefactor held constant) can be a negative cue of social valuation: for instance even if I have a low WTR toward you, I will still help you if the benefit to you is particularly large and my WTR is non-negative. On the other hand, large benefits may signal that the benefactor put some effort into ensuring the effectiveness of their good act, which would reveal a high valuation.

[14]Sell (2005, p.232) writes "Intentionality remains to be fully explored and mapped computationally such that it can be modeled without intuitive references.". Tamsin German and I have started working in that direction. We argue that people consider an agent to have done X intentionally if the agent's attitude toward X (i.e. how much the agent wants or does not want X to happen) caused X to happen (Quillien & German, under review; see Quillien, 2020b for a computational model of the evolved concept of cause). If we accept this account, then the intentional status of the action makes it diagnostic because it reveals that the benefit delivery was caused by the benefactor's valuation of the recipient.

are an important input to gratitude.

More recently, Monroe (2020) tested whether the moral emotion of elevation (Fessler & Haley, 2003) is regulated by inferences about social valuation. Using a design similar to Sell (2005), in a first study she found that, holding benefit delivery constant, witnessing a good deed triggered higher elevation when the benefactor paid a higher cost. However, this effect was absent in a follow-up replication attempt using a different vignette and a different measurement strategy.

Emotional adaptations coevolve with conceptual systems (Barrett, Cosmides, & Tooby, 2010; Delton & Sell, 2014). If inferences about WTPs matters to social perception, it is likely that people spontaneously categorize others according to their WTPs. This hypothesis was tested by Delton & Robertson (2012) using the "who-said-what?" task, an implicit measure of categorization. They asked participants to read about fictitious people (henceforth, 'targets') stranded on a desert island. Participants read sentences depicting the people foraging for food. Some of the sentences depicted a target as paying large costs to get food for the group, while others depicted another target as paying small costs to get a similar amount of food. Costs incurred were diagnostic of the person's willingness to contribute to the group, and therefore were a cue to WTPs. Delton & Robertson (2012) found that participants spontaneously categorized targets according to the costs they incurred. By contrast, they failed to spontaneously categorize targets by costs in similar scenarios where the targets were foraging for themselves instead of foraging for the group. They also failed to categorize targets according to the size of the benefits they provided to the group when variation in benefits was due to luck and therefore not diagnostic about WTPs. This is evidence that WTPs provide a fundamental dimension along which people spontaneously categorize others.

Most of the evidence reviewed so far does not distinguish between inferences about the WTPs of others toward the self in particular (i.e. Bob's inference about Alice's

WTPs toward Bob) and inferences about a general other-regarding disposition (Alice's WTPs toward everybody else). People care about self-specific WTPs in their choice of partner (see Lukaszweski & Roney, 2010), and there is evidence that they make inferences accordingly. Sell (2005) showed that if Alice plays a mean prank on Bob, Bob is angrier if he knows that Alice intentionally targeted him in particular. Sznycer (2010) asked whether people can estimate (based on their own experience) their friends' WTR toward them, above and beyond their friends' overall generosity. He recruited friend dyads and measured, for each participant, the participant's WTR toward their friend, their WTR toward an acquaintance, and their estimate of their friend's WTR toward them. Across dyads, participants' estimates of their friends' WTR toward them were correlated with their friends' WTR toward them, even controlling for their friends' WTR toward an acquaintance (partial r=.27); by contrast, there was no significant zero-order correlation between a participant's estimation of their friends' WTR and the friends' WTR toward an acquaintance.

Although less direct, convergent evidence for the role of social valuation inference in human social cognition comes from a large literature on the 'dimensions' of social perception. Social psychologists have long recognized that we evaluate others not via simple associative learning, but by assuming that behavior is a joint product of a person's traits and the relevant context. Information about the target's behavior, in conjunction with information about the context, leads people to draw causal inferences about the target's traits (Heider, 1958; Kelley, 1973; Ajzen & Fishbein, 1975). Later research in social psychology found that humans everywhere interpret behavior and form impressions of others primarily along two dimensions: 'warmth' and 'competence' (Fiske, Cuddy & Glick, 2007). 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.

While both warmth and competence information matter to impression formation, warmth information usually has a larger influence (Fiske, Cuddy & Glick, 2007; see also Eisenbruch & Krasnow, 2019). 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 et al., 1998). People are also disproportionately interested in gathering information about warmth-related traits: when asked which traits they most would 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).

In sum, evidence from a variety of tasks, using a large array of dependent measures (impression ratings, emotional reports, partner choice decisions, spontaneous categorization, BOLD signal, information-gathering decisions, monetary allocations, etc), support the hypothesis that humans make inferences about social valuation, and that these play an important role in social cognition.

However, this body of research is relatively silent about how we make such inferences (beyond the fact that people use the relevant cues). Here, I will argue that an adaptationist approach not only predicts that people make social valuation inferences, it can tell us how they do it. Starting from a task analysis of the inference problem, one can generate quantitative predictions about the form of people's inferences.

1.4 A task analysis of social valuation inference

1.4.1 Reasoning under uncertainty

In its general form, the inference problem we are interested in is the following. Given observable evidence (e.g. Alice's words and deeds), how should Bob update his representation of the WTPs of Alice towards him? I will consider this general problem at the computational level of analysis (Marr, 1982); in other words, how would one design a machine that makes such inferences optimally?

Bob is facing a problem of inference under uncertainty. A given observation does not logically entail the veracity of a single hypothesis to the exclusion of all others. For instance, the observation that Alice did not share her cake with Bob is consistent with many hypotheses about how much she values Bob: maybe she actively wishes for Bob to starve to death; maybe she likes him but cares about herself more; maybe she mistakenly thought he was on a diet. Bob cannot determine the exact value of the weight that Alice puts on his welfare on the basis of this one observation, so the best he can do is narrow down his probabilistic estimate of that weight.

Sound reasoning under uncertainty has been extensively studied by mathematicians, philosophers, and computer scientists. It can be shown that any system for reasoning under uncertainty that satisfies certain elementary desiderata (e.g. consistency) has to be isomorphic to probability theory (Cox, 1961; Jaynes, 2003). That is, an ideal reasoner should assign real numbers to hypotheses, representing her degree of belief in that hypothesis. She should also update her belief in a given hypothesis, given new evidence, by using the laws of probability. Reasoners who assign beliefs in a manner inconsistent with probability theory expose themselves to 'Dutch books': one can construct a bet where they are guaranteed to lose money, no matter the outcome of the event that the bet is about (de Finetti, 1931). The interpretation of probability theory as a reasoning system

is often called 'Bayesianism', in reference to Bayes' rule (Bayes, 1763; Laplace, 1812), a simple theorem of probability theory which implies a specification of the rational way of updating one's beliefs given new information. Since most of the inference problems solved by evolved organisms involve inference under uncertainty, Bayesianism is highly relevant to efforts to reverse-engineer the evolved mind (Cosmides & Tooby, 1996; Pietraszewski & Wertz, 2011; Barrett, 2014).

Therefore, one computational-level requirement for social valuation inference is consistency with probability theory.

A brief look at Bayes' rule makes evident that inference under uncertainty also requires domain-specific background knowledge. Bayes' rule says that the rational way to update one's belief in hypothesis H, given the observation of new data D, is via the formula:

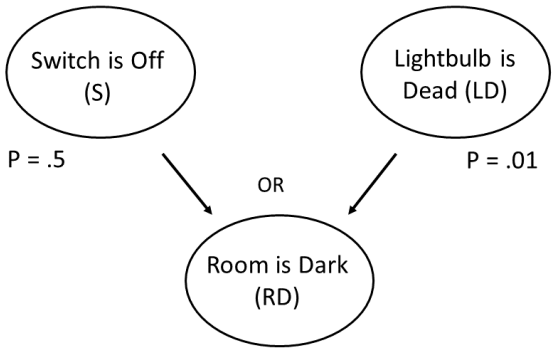
$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

The $P(D|H)$ term is called a likelihood: it is the probability of observing the data assuming the hypothesis is true. To evaluate this conditional probability, the reasoner needs background knowledge that specifies what one would expect the world to be like if hypothesis H held true. Very often, this background knowledge will take the form of a causal model of the world. Here is a very simple example to make this point more concrete. Suppose I observe that the light in my room is off, and I want to infer the plausibility of the hypothesis "the ligthtbulb is dead". Let the variable RD denote whether my room is dark (the data), and LD denote whether my lightbulb is dead (the hypothesis). To infer whether the lightbulb is dead, I must use Bayes' rule:

$$P(LD|RD) = \frac{P(RD|LD)P(LD)}{P(RD)}$$

In order to evaluate the probabilities on the equation's right-hand side, I need to have

a causal model of the situation. Here, a plausible causal model would be the following:



This diagram says that the room will be dark if the switch is off or the lightbulb is dead. It also says that the prior probability of the switch being off is .5, while the prior probability of the lightbulb being dead is .01. It is easy to compute, then, that $P(RD) = P(LD) + P(S)$–$P(RD)P(S) = .5 + .01 - .5 * .01 = .505$.

The likelihood term, $P(RD|LD)$ can be read off from the causal model: here it is simply 1, since the lightbulb's death is sufficient for the room being dark. Therefore Bayes' rule gives us $P(LD|RD) = \frac{1*P(LD)}{P(RD)} = .01/.505 \approx .02$. This formalizes the commonsense intuition that, despite the room being dark, it is unlikely that the lightbulb is dead, since the switch being off provides a more likely explanation.

In this example, the necessary background knowledge comes from learned information about how lamps work. But in general, one expects inference mechanisms to come equipped with reliably-developing background knowledge about recurrent features of our species' ancestral environment (Boyer & Barrett, 2015; Quillien, 2018). First, genetically built-in inductive priors speed up learning, providing a non-trivial adaptive advantage; second, the prior and likelihood terms that, according to Bayesianism, are prerequisites for learning, need to come from somewhere: although they can themselves be learned (Kemp, Perfors & Tenenbaum, 2007), the learning process must reach an unlearned bedrock at some point (Barrett, 2014).

In sum, very general computational considerations lead one to predict that human social valuation inference involves a Bayesian updating process, and relies on domain-specific causal knowledge.

Some readers may find that this computational-level analysis, though interesting, is simply irrelevant to understanding how the mind solves the problem. Haven't humans been shown over and over to fail to conform to normative standards of probabilistic reasoning? In the next section I try to reconcile the current prediction with these findings.

1.4.2 When do we expect humans to conform to normative standards of probabilistic reasoning?

There is a long tradition in cognitive psychology, social psychology, and behavioral economics, to compare human probabilistic reasoning to normative standards. The results are often unflattering for our species. Along with failures to adhere to normative standards in other domains, these results have contributed to the oft-popularized idea that humans are 'irrational' (Kahneman, 2011; Marcus, 2008; Ariely, 2008).

For instance, people tend to ignore relevant base-rate information when making judgments under uncertainty. When computing the probability that a patient with a positive test actually has the disease, they disregard the base rate of the disease, massively underestimating the probability of a false positive (Casscells, Schoenberger & Graboys, 1978); when asked, on the basis of imperfect witness testimony, to evaluate the probability that an accident was caused by a Green taxi, they disregard information about the market share of the Green taxi company (Bar-Hillel, 1980). People also make mistakes that violate basic set-theoretic principles underlying probability theory: they sometimes judge the probability of A & B as strictly higher than the probability of A (Tversky & Kahneman, 1983; Ludwin-Peery, Bramley, Davis & Gureckis, 2020).

How might one account for these findings in the context of the computational considerations I laid out in the previous section?

Here it is helpful to think about the cognitive architecture underlying inference. A naïve model of this architecture would be to posit the existence of a 'Bayes box' somewhere in the brain to which every problem of inference under uncertainty that the mind has to solve is automatically routed. The importance of domain-specific background knowledge for inference militates against that proposal. Instead, one expects the existence of a variety of dedicated inference systems in the mind, each equipped with the background knowledge appropriate to its domain (Barrett, 2014)[15]. Relatedly, the mind has to solve a difficult routing problem: given an inference problem, the mind has to detect that it is an inference problem, and route it to the appropriate inference system (Barrett, 2005).

This routing problem is far from trivial, and it is plausibly one of the main sources of 'irrationality' in human probabilistic inference. One expects that people will be unable to conform to the normative standard when their mind either (a) fails to classify an input as an inference problem, (b) routes the input to an inappropriate inference system, (c) fails to correctly 'translate' the input.

There is strong evidence that something like this must be going on with base-rate neglect in medical-diagnosis problems. Cosmides & Tooby (1996; see also Gigerenzer & Hoffrage, 1995) showed that the input format of the problem has extremely large

[15]There are other reasons to expect modularity when it comes to inference. For instance, some inference problems need to be solved faster than others: "is this a snake or a broken branch in front of me?" needs to be solved faster than "is it going to rain later today?". One expects inference systems that deal with such time-sensitive information to trade-off accuracy for speed. Also, inference systems differ in the extent to which the inferences they generate need to be explainable to others. For instance, our visual system automatically infers the reflectance (i.e. the 'objective' color) of objects based on cues such as luminance, background illumination, etc (Shepard, 1992), but we rarely need to explain to others why we think that a given object is red. By contrast, being able to explicitly justify one's conclusions is important in other domains, for instance those related to social interaction (Mercier & Sperber, 2017; Mahr & Csibra, 2017; see also the argument made about anger a few sections earlier).

effects on whether people give the correct answer. When given the problem in terms of probabilities (e.g.: 0.1% of people have the disease, the test has a false positive rate of 5%, what is the probability that someone who tests positive actually has the disease?), participants gave massive overestimates (the modal answer being 95%), but when reading the problem presented in terms of frequencies (e.g.: out of 1000 people, 1 has the disease, and 50 healthy people will test positive), most participants gave the correct answer (about 2% of people who test positive actually have the disease). Therefore, base-rate neglect in the medical diagnosis problem (when it happens) is not due to a blanket incapacity of the human mind to correctly carry out Bayesian computations.

More likely, it is due to a failure to correctly translate the input into the right mentalese equivalent, and/or a failure to route the input to the relevant inference system. Krynski & Tenenbaum (2007) suggest that the issue arises because the mind fails to translate the information that participants are given into an appropriate causal model. Specifically, because participants have not been given plausible causes for the false positive tests, they do not include the corresponding nodes in their causal model. As a result, they find it difficult to 'explain away' the positive test with an alternative to the hypothesis that the patient has the disease. As support for their claim, Krynski & Tenenbaum find that in a version of the medical diagnosis test (with a probability format) which specifies that positive tests can also be caused by a benign cyst, most participants give the correct answer[16].

Incorrect human reasoning is also expected when the input is routed to the correct inference system, but is mistranslated such that the mind correctly answers a different question than the one asked by the experimenter. Researchers have claimed that humans

[16]It is notable that a frequentist format allows people to perform correct Bayesian computations despite no explicit representation of an alternative cause for the positive test. This suggests the existence of at least two distinct potential mechanisms via which the mind can successfully solve a problem like the medical diagnosis scenario.

do not understand randomness, because they view a series of coin tosses like 'HHHHHH' as less likely to occur than 'HHTHTT' (both series are equally likely). This answer is incorrect if we assume that the mind is trying to compute $P(`HHHHHH'|coin\,is\,fair)$. However, further research has shown that people's judgments about randomness are actually close to optimal if we assume that their mind is trying to estimate the converse conditional probability, that is $P(coin\,is\,fair|`HHHHHH')$. A series of six Heads warrants skepticism about whether the coin is truly random, more so than a series that contains both Heads and Tails (Griffiths & Tenenbaum, 2001).

The evidence reviewed so far shows that people's reasoning about explicit problems often exhibits the fingerprints of Bayesian inference. But by far the largest source of evidence for Bayesian belief updating comes from studies of implicit inference problems, of the kinds involved in perception. Scores of phenomena, involving object recognition (Knill & Richards, 1996; Kersten, Mamassian & Yuille, 2004), multi-modal cue integration (Ernst & Banks, 2002), motion perception (Weiss, Simoncelli & Adelson, 2002), and face processing (Peterson & Eckstein, 2012) show that human perception often conforms to Bayesian standards.

It is also noteworthy that information processing in organisms with a considerably simpler nervous system than humans is often well-described by normative theories of statistical inference. For example, bumblebees optimally learn the reward structure of their environment when foraging (Real, 1991; Biernaskie et al., 2009).

In sum, what does the available evidence suggest are the sources of human difficulties in probabilistic reasoning?

Processing limitations are an unlikely explanation. From a computational complexity perspective, tasks such as the medical diagnosis problem are trivial; and even though Bayesian inference can be intractable in more challenging settings, well-known techniques such as Markov Chain Monte Carlo provide tractable approximations, and there is evi-

dence that the human mind may employ similar techniques (Vul, Hanus & Kanwisher, 2009; Gershman, Vul & Tenenbaum, 2009; Lieder & Griffiths, 2020). From a comparative perspective, if bumblebees have the processing power to use Bayes' rule, it is likely that humans do as well. The hypothesis that humans simply did not evolve mechanisms capable of Bayesian inference is refuted by data showing that the mind is actually a good statistician in many tasks, implicit and explicit. The most likely explanation of failures to uphold good rules of statistical reasoning is that the input (e.g. a verbal description of a problem) is not adequately carried in the right format and/or to the right place in the participant's mind.

Thus, an evolutionarily analysis can explain when human statistical inference is and isn't successful by considering for which content natural selection was plausibly able to design input analyzers that successfully route the input to the appropriate inference system. For instance, in the context of perception, there has been a strong selection pressure for the design of systems that automatically pick up on the relevant cues and integrate them optimally to form an accurate representation of one's immediate surroundings. In the context of explicit probabilistic problems, mathematical tools such as percentages are evolutionary novel, so it less likely that we would have built-in equipment that can adequately process them.

So where is social valuation inference likely to belong? Is it more like perception, or more like reasoning about percentages? Earlier, I have given arguments for why selection is likely to have favored adaptations for social valuation inference. This leads us to expect that natural selection was able to design the necessary infrastructure to tag an input as relevant to social valuation inference, and route it to the relevant inference system. In consequence I predict that people make approximately Bayesian inferences in this context.

1.4.3 An ideal observer model of social valuation inference in a simple task

A completely general formal model of the way humans infer the WTPs of others is beyond the scope of the present work. Given the complexity of the machinery that regulates welfare-tradeoffs, one expects that the machinery for social valuation inference is correspondingly complex. Instead I will study the inference that people make in a very simple inference task. In this task, it is easy to derive an 'ideal observer model': a mathematical model of the optimal inference that one can make given the available data.

The task is the following: Alice and Bob play a few rounds of the WTT, with Alice as dictator and Bob as recipient. How should Bob update his estimate of Alice's WTR towards him, given his observations of Alice's decisions?

As a reminder, the WTT (Welfare Tradeoff Task) 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 ideal observer model relies on a causal model that represents how Alice's decisions are caused by the specific payoffs of the current round and Alice's WTR toward Bob.

I constructed this causal model on the basis of empirical data about how people typically behave in the WTT when playing as dictators (Delton, 2010). Alice plays the WTT so as to maximize her expected utility, given by Eq. 1:

$$U_{alice} = \pi_{alice} + WTR_{alice->bob} * \pi_{bob} \quad \text{(Eq. 1)}$$

The decision rule that follows from this utility function is that Alice allocates the money to Bob if

$$WTR_{alice->bob} > \frac{\pi_{alice}}{\pi_{bob}}$$

We also assume that Alice observes a noisy value of the payoffs in each trial. Specifically, for each trial she observes a noisy value of $\phi = \frac{\pi_{alice}}{\pi_{bob}}$, with some added noise ϵ that is drawn from a normal distribution with mean 0 and variance σ^2_{phi}. 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 (Fisman, Kariv & Markovits, 2007; Delton, 2010).

Note that this way of modeling noise in Alice's decisions is somewhat arbitrary. Alternatively, I could have assumed that Alice has a perfectly precise mental representation of the payoffs for the current trial, but her WTR is itself noisy (for instance, every time she makes a decision she draws her WTR as a sample from a probability distribution centered on her 'true WTR' towards Bob). Or I could have assumed that Alice has perfectly precise representations of the payoffs, and of her WTR towards Bob, but that she chooses an action (Give vs Take) using something like a softmax decision function, choosing a given action with probability proportional to its expected utility. I am not aware of any existing data that could discriminate among these alternatives. However, for modeling purposes this does not matter, because any combination of these assump-

tions would make the same behavioral predictions (e.g. Alice is more likely to make a decision inconsistent with her WTR when the ratio $\frac{\pi_{alice}}{\pi_{bob}}$ is very close to her WTR).

This causal model makes it possible to compute a likelihood term, $P(decision|WTR, \phi)$, which expresses 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:

$$P(`Give'|WTR, \phi) = P(WTR > \phi + \epsilon)$$

$$P(`Take'|WTR, \phi) = 1–P(`Give'|WTR, \phi)$$

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 $P(WTR)$: 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 $P(decision|WTR, \phi)$ for different WTRs, where each possible WTR is weighted according to its probability. Formally, we write this:

$$P(decision|\phi) = \int P(decision|WTR, \phi)P(WTR)\, dWTR$$

We now have all the necessary pieces to implement Bayes' rule. 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 the following equation:

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

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

Although the ideal observer's belief about a partner's WTR is a probability distribution, it is often more convenient and intuitive to consider it as a single number. In analyzing results, when I refer to the WTR that the observer infers a partner to have, I am using the median of this distribution.

Algorithmically, I used grid approximation to implement the ideal observer. The R code for the implementation is available at the Open Science Framework[17].

1.4.4 The current studies

The studies presented in this dissertation test whether the ideal observer model I just described is a good fit for actual human inference. Chapter 2 presents a simple test of the model, in two studies where participants were shown a few WTT choices made by an agent, and were asked to predict what this agent did in other trials of the WTT. I compare their predictions to those made by the ideal observer. The two studies in chapter 2 also present a strong test of the hypothesis that social valuation inference is an input to anger and gratitude. In chapter 3, I test whether people are spontaneously curious about the pieces of evidence that would provide the most information to the ideal observer.

[17] https://osf.io/bf6s4/?view_only=6b47266a55b847bab14a13f4d426292d

Chapter 2

A simple test of the rationality of human social valuation inference

2.1 Introduction

Research reviewed in the previous chapter suggests that people make inferences about the WTPs of others by using the relevant cues (costs, benefits, intention, etc). However, this research did not investigate whether people end up forming accurate estimates of the WTPs of others. It provides evidence that people are sensitive to relative differences in WTPs between two people; for instance participants in Lim (2012) were sensitive to the fact that one of their partners had a higher WTR than the other. However, these studies do not show that people make the kind of inferences that would allow them to make quantitative predictions about the behavior of the targets.

It would be possible to design a study that probes whether people make such quantitative inferences by, e.g., making a slight change to the design used by Lim (2012). Allow the participant to observe many choices that a (sham) partner makes in the WTT, such that an ideal observer would infer with a high degree of confidence that the partner has a

WTR of (say) .7. Then ask the participant to predict whether the partner would Give or Take in other trials. For instance, if the partner has a choice between $10 for herself and $30 for the participant, a belief that the partner has a WTR of .7 toward the participant would lead one to predict that the partner will Give in this trial.

However, such a design would not be able to uncover particularly strong evidence that participants are updating their belief in an approximately Bayesian fashion. There are many sub-optimal learning algorithms which, given sufficient data, will eventually converge to the truth. The fingerprints of Bayesian updating are more evident when the agent needs to makes inferences from sparse data. Therefore, a particularly strong test of people's inferential abilities is to ask them to make predictions after having only seen a few WTT decisions made by their partner. In this context, observers can only make probabilistic predictions (e.g. "I think that Alice will Give in this trial with 62% probability"), but it is possible to quantify the correct probability that one should assign to a given outcome. This is the approach I use in the current studies.

Some existing studies have sought to formally model the learning process by which a participant infers something like the WTR of a target, using Bayesian or reinforcement learning models (e.g. Xiang, Lohrenz, & Montague, 2013; Siegel et al., 2018; Hackel et al., 2015, 2020). However, these studies were not specifically designed to test the fit between these models and human behavior; for instance, they allowed a long learning process spanning many trials.

The studies reported in this chapter also have a second goal: to test whether inferences about social valuation are an input to computational systems regulating anger and gratitude.

According to the recalibrational theory of anger, anger is triggered by decisions that elicit inferences that the actor puts a low weight on one's welfare. Existing studies have tested this hypothesis by using scenarios with roughly the following structure:

-Alice makes a decision that makes her win $2 but makes you lose $10

-Claire makes a decision that makes her win $30 but makes you lose $10

People are typically angrier at Alice than Claire – this intuition appears to hold cross-culturally (Sell et al., 2017). Under the welfare valuation inference hypothesis, this is because the two decisions elicit different inferences about the weight that the agent puts on one's welfare.

However, such findings are open to alternative interpretations. Emotions might be driven by simple heuristics rather than probabilistic inferences per se. A heuristic such as "be angrier at people who get small benefits when they harm you" would also predict that people would be angrier at Alice than Claire. In order to provide a stronger test of the inference hypothesis, I probed participants' emotion judgments toward agents who (i) inflict the same total cost to the participant, (ii) reap the same total benefit from their offenses, yet (iii) elicit different welfare valuation inferences in an ideal observer. For instance, imagine that Alice and Claire make two decisions each:

-Alice makes Bob lose $10 in order to get $1. Later, she makes Bob lose $1 in order to get $20.

-Claire makes Bob lose $10 in order to get $10. Later, she makes Bob lose $1 in order to get $11.

In both cases, the agent gets a total benefit of $21 and inflicts a total cost of $11 on Bob. Yet Alice's decisions – in particular, her first decision – are much more informative about how much she values Bob. Therefore, a Bayesian ideal observer will estimate from these data that Alice probably values Bob less than Claire does. This predicts that

participants will be angrier at Alice. By contrast, a simple heuristic view, based on a simple tally of costs and benefits, predicts equally intense anger toward both agents.

I applied the same strategy to assess judgments of gratitude, by designing players who generated the same amount of benefits to the participants, incurred the same total cost for doing so, yet elicited different inferences in the ideal observer.

2.2 Methods

2.2.1 Task

Welfare-tradeoff task

Participants were first familiarized with a simple economic game, described to them as a "money allocation task". The game was the Welfare-Tradeoff Task (WTT – Delton, 2010; Delton & Robertson, 2016; see Chapter 1 for details). To familiarize themselves with the WTT, participants played four rounds of a pretend version of the game in the role of dictator, while being asked to imagine that the receiver was one of their acquaintances. Throughout the study, no money was involved, but participants were asked to imagine that they were playing for real money.

Prediction task

In the main task, participants played the WTT in the role of the receiver. The dictators they played with were fake partners generated by the computer. Participants were aware of this – no deception was involved at any point in the study. I asked participants to imagine each partner as one of their acquaintances – a different acquaintance for each partner.

Each participant played the WTT with 10 partners in total – partners always played

as dictators, while participants played as receivers. For each partner, they first saw that player make two decisions; after seeing the two decisions, the participant was asked to rate how angry and how grateful they felt toward their partner, on two 1-7 likert scales.

They were then shown 5 other trials of the WTT that their partner had played; for each of them, they were asked to predict, using a slider scale from 0% to 100% likely, the probability that their partner chose to allocate the money to the participant on that trial. I counterbalanced the framing of the question such that half the participants were actually asked to rate the probability that the partner would allocate the money to themselves, and I reverse-coded the ratings for these participants. Each trial was displayed on a separate page; each page also displayed, as a reminder, the two decisions that the participants had initially observed. I did not give feedback to the participants' predictions.

Partners were presented in random order. Among the 10 partners each participant played with, 5 were "selfish" partners who always allocated the money to themselves in the two trials observed by the participant, while 5 were "generous" partners who always allocated the money to the participant. Table 2.1 shows the decisions made by each partner, and the potential payoffs for each trial. The order in which the decisions made by a partner were presented was counterbalanced across participants.

Table 2.2 shows the potential payoffs for the five trials for which participants had to make predictions. These 5 trials were identical for all partners, but were presented in a randomized order within a partner.

I designed the partners such that among the selfish partners, over the two decisions that the participant saw that partner make, each partner received the same total payoff, and inflicted the same total opportunity cost on the participant. Similarly, among the generous partners, each partner incurred the same total opportunity cost, and allocated

the same total amount of money to the participant[1]. On the other hand, even among partners of the same type (e.g. selfish partners), different partners elicited a wide range of different inferences in the ideal observer about the partner's WTR toward the participant.

For each partner, I computed the WTR estimate that the ideal observer would infer that partner to have toward the participant. The ideal observer initially had the same prior belief for the WTR of each partner, and then updated that belief as a function of the two decisions observed by the participant.

For instance, partner A makes two decisions. In one of them, he gives \$29 instead of taking \$29. In the second, he gives \$96 instead of losing \$15 (see table 2.1). The second decision is not very informative about his WTR, so it does not trigger a large update in the ideal observer's belief. By contrast, when observing the first decision, the ideal observer can draw the inference that partner A's WTR is probably[2] above 1 (since with a WTR below 1 he would have chosen to take \$29 instead of giving \$29). As a result, the ideal observer shifts the probability mass of its belief toward WTR values higher than 1. Here, given the prior I use, the updated probability distribution has a median of WTR = 1.53.

[1]Due to a simple addition error when designing the study, partner D gave a total benefit of \$123 instead of \$125 to the participant.

[2]I say "probably" instead of "certainly" because the ideal observer assumes that the dictator's choices are not completely deterministic.

Partner	$\pi_{partner}$	$\pi_{participant}$	decision	ideal-observer-inferred WTR
A	29	29	Give	
	-15	96	Give	1.53
B	24	33	Give	
	-10	92	Give	1.35
C	15	27	Give	
	-1	98	Give	1.24
D	12	31	Give	
	2	92	Give	1.14
E	5	35	Give	
	9	90	Give	1.04
F	50	29	Take	
	10	6	Take	0.36
G	26	33	Take	
	34	2	Take	0.01
H	16	27	Take	
	44	8	Take	-0.1
I	12	31	Take	
	48	4	Take	-0.24
J	1	35	Take	
	59	0	Take	-0.49

Table 2.1: Decisions made by each partner, along with the WTR that the ideal observer model inferred the partner to have after the model observed both decisions. The order in which the decisions made by a partner were presented was counterbalanced across participants.

$\pi_{partner}$	$\pi_{participant}$	ϕ
1.5	30	.05
7.5	30	.25
16.5	30	.55
27	30	.9
39	30	1.3

Table 2.2: Potential payoffs for the partner and the participant, for the five prediction trials. $\phi = \frac{\pi_{partner}}{\pi_{participant}}$

2.2.2 Procedure

After signing a consent form, participants were explained the structure of the WTT, and then played four rounds as a dictator to familiarize themselves with the task. Then they completed the prediction task. Finally, they completed a few demographic questions, and were thanked for their participation.

2.2.3 Parametrization of the ideal observer

For each prediction that participants had to make, I computed the prediction made by the ideal observer (see Chapter 1 for description of the model).

The ideal observer must be equipped with a prior. This corresponds to the observer's baseline expectation about the WTR of a partner for which the observer has no information. In study 1, I tested a parametrization that relies on the hypothesis that people have a good representation of the statistical regularities of the environment, and that therefore their priors reflect the actual distribution of WTRs in the population. In order to establish this prior, I used empirical data from a larger study (Sznycer et al., unpublished data) where participants (N = 479, recruited on Amazon MTurk) played as dictators in the WTT. I inferred the statistical distribution of WTRs in this sample (see Appendix B for details), and used this distribution as the prior for the ideal observer.

Finally, the causal model used by the ideal observer features a parameter σ_ϕ, quantifying the amount of noise that goes into people's welfare-tradeoff decisions (see Chapter 1). I set this parameter's value by inferring the median value of σ_ϕ in the same MTurk sample I used to derive the first parametrization of the prior (see Appendix A).

The ideal observer model, as well as the design of studies 1 and 2, were pre-registered[3]; and the studies were approved by the Institutional Review Board at UCSB. The data,

[3]https://osf.io/y8hks/?view_only=9948bf341a3d4c5d8df4aaf0d9baabd4

and the R code for the computational model, data analysis and figures are available at the Open Science Framework[4]. For some statistical tests I use linear mixed models; when I do so I z-score the predictor and outcome variables, so that b coefficients can be interpreted as effect sizes.

2.3 Study 2.1

2.3.1 Participants

I recruited 100 US residents (40 female, mean age: 34.11) from Amazon Mechanical-Turk. I excluded 37 participants who failed an attention check, yielding a final sample of 63 (26 female, mean age: 34.86). I chose this sample size because it seemed very large, given the large number of trials per participant and the within-subjects nature of the main tests.

2.3.2 Results

Do human predictions match ideal observer predictions?

Yes.

Participants had to make 50 predictions (5 predictions per partners, for 10 partners). For each such prediction trial, I computed the average probability that participants assigned to "partner Gives" in that trial, and I also computed the prediction made by the ideal observer.

The item-level correlation between the average human prediction for a given trial and the model prediction for that trial was $r(48) = .86$, $p < .001$. Human predictions also correlated with model predictions when analyzed at the individual level: the median

[4]https://osf.io/bf6s4/?view_only=6b47266a55b847bab14a13f4d426292d

correlation between an individual's predictions and the model predictions (across trials) was r(48) = .75; inter-quartile range: .66 to .84.

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

Maybe participants did not engage in social valuation inferences, but made predictions by simply computing the sums of the benefits and opportunity costs associated with a partner's two decisions in the observation trials, and/or making less optimistic predictions in prediction trials when $\pi_{dictator}$ (the opportunity cost of giving) was large.

In order to rule out that possibility, I computed the association between the WTR that the ideal observer inferred the partner to have toward the participant, and the average human predictions for that partner. I did so while statistically controlling for a dummy variable indicating whether a partner was "selfish"or "generous". Recall that here I call "selfish" a partner who Takes the money for herself in both decisions that the participant observes, and I call "generous" a partner who Gives the money in both decisions that the participant observes. Henceforth, I refer to this dummy variable as "material payoffs". This is because all 5 selfish partners make decisions with the same aggregate material consequences, in terms of benefits gained and opportunity costs inflicted (and similarly for all 5 generous partners). If participants did not make WTR inferences, they would make the same predictions for all 5 "generous" partners, and they would make the same predictions for all 5 "selfish" partners.

A linear mixed model with partner's type and inferred WTR as predictors, random slopes and random intercepts, and participant as a random effect, shows that controlling for material payoffs, the WTR inferred by the ideal observer remains positively associated with human predictions, b = .09, p = .006, suggesting that participants did make social valuation inferences.

Does inferred WTR predict anger and gratitude?

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

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

2.4 Study 2.2

Study 2 is a replication and extension of Study 1. In study 1, I calibrated the ideal observer with a prior that reflected the distribution of WTRs in a sample of MTurk participants playing the WTT as dictators. This calibration strategy is based on the hypothesis that participants have, over the course of their lives, learned a good representation of the statistical distribution of the WTRs of their acquaintances. There are potential issues with this strategy, however. First, I computed the distribution of WTRs among a sample of MTurk workers, who might not be representative of the typical acquaintances that our participants typically interact with. Second, even assuming that participants do have a good representation of the Welfare Tradeoff Parameters of others, they may be uncertain about how this would translate in the context of an artificial laboratory experiment like the WTT.

Therefore, in study 2, I attempted to measure participants' priors more directly. The study was similar to study 1, with the exception of an additional phase at the beginning of the study. In this preliminary phase, participants were asked to make predictions about dictators for whom they had no information about past behavior in the Welfare Trade-off Task. I used their predictions in this phase to estimate the prior beliefs that participants

have about the distribution of WTR among their acquaintances. I then used these estimates to determine the prior of the ideal observer. Study 2 also used an undergraduate sample instead of an online sample, because I thought an undergraduate sample would yield more precise individual-level data, given the more controlled environment of the laboratory.

After completing the WTT familiarization phase, but before the prediction task, participants were asked to complete a variant of the prediction task where they had to predict the behavior of 20 different interaction partners for whom they had not observed any prior decision. They made one prediction per partner, in trials of the WTT with $\pi_{participant}$ = \$30 and $\pi_{partner}$ ranging from \$3 to \$60 in \$3 increments (trials were presented in randomized order). I asked participants to imagine each partner as one of their acquaintances – a different acquaintance for each partner. Using these 20 predictions, one can infer the prior that the participant has about the WTR of his average acquaintance. I did so for each participant. I averaged these priors to generate a prior for the ideal observer (see Appendix C for details). In most of the analyses that follow, I use this prior generated by averaging the priors of all participants. However, in section 2.8 I use participant-specific ideal observers that are equipped with the prior inferred for an individual participant.

2.4.1 Participants

I recruited 100 participants (72 female, 1 other, mean age: 18.8) from the undergraduate psychology subject pool at a university in California. Participants completed the study on a desktop computer while seated in a semi-private cubicle. One participant failed to complete the study because of computer error. I excluded from analysis 32 participants who failed either a probability comprehension check (4 participants) and/or

an attention check (29 participants) yielding a final sample of 67 participants (48 female, 1 other, mean age: 18.8).

2.4.2 Results

Do participants' priors differ from the prior used in Study 1?

Yes. The best-fitting prior had a mean WTR of .55 and standard deviation of 1.01. By contrast the prior used in study 1 had a mean WTR of .22 and standard deviation .39. This suggests that the uncertainty that participants had about the potential WTR of an acquaintance playing the WTT was higher than that suggested by simply computing the variance of WTRs in an MTurk sample. Henceforth all results I report will use the ideal observer calibrated with the new prior.

Do human predictions match ideal observer predictions?

Yes. The item-level correlation between the average human prediction for a given trial and the ideal observer's prediction for that trial was very large, $r(48) = .988$, $p < .001$; see Figure 2.1.

Can this result be explained by simple heuristics?

No. Controlling for material payoffs, the WTR inferred by the ideal observer was positively associated with human predictions, $b = .52$, $p < .001$; (linear mixed model with random slopes and random intercepts, material payoffs and inferred WTR as fixed effects, and participant as a random effect).

Does the WTR inferred by the ideal observer predict anger and gratitude?

Yes. Figures 2.2 2.3 display Anger and Gratitude ratings of participants for each partner, as a function of the WTR inferred by the ideal observer for that partner. Inferred WTR was a negative predictor of Anger, $b = -.65$, $p < .001$, and a positive predictor of Gratitude, $b = .86$, $p < .001$ (linear mixed models with random slopes and random

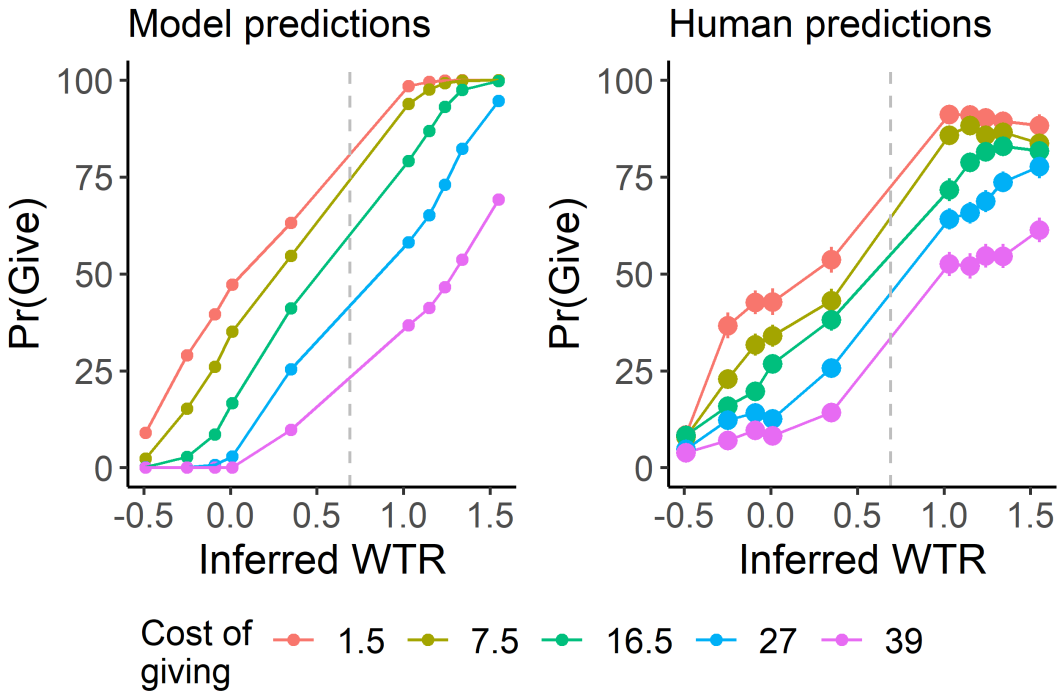


Figure 2.1: Predictions made by the ideal observer (left) and average predictions made by human participants (right), in Study 2. Each dot represents one trial. In both panels, the x-axis represents the WTR that the ideal observer inferred the partner to have toward the participant. "Cost of giving" is the potential payoff (in USD) for the dictator in that trial. Error bars represent standard errors of the mean. Within each panel, selfish partners are at the left of the dashed line, while generous partners are at the right of the dashed line.

intercepts, inferred WTR as fixed effect, and participant as a random effect).

When controlling for material payoffs, inferred WTR remained a significant predictor of Anger, b = -.73, $p < .001$, and Gratitude, b = .10, p = .02, although people's Gratitude ratings were mostly driven by material payoffs.

2.5 Reanalysis of study 1

It is possible that participants in study 1 (although recruited from a different pool) have priors that are relatively similar to those of participants in study 2 (for instance, perhaps most people living in the US have relatively similar baseline expectations about

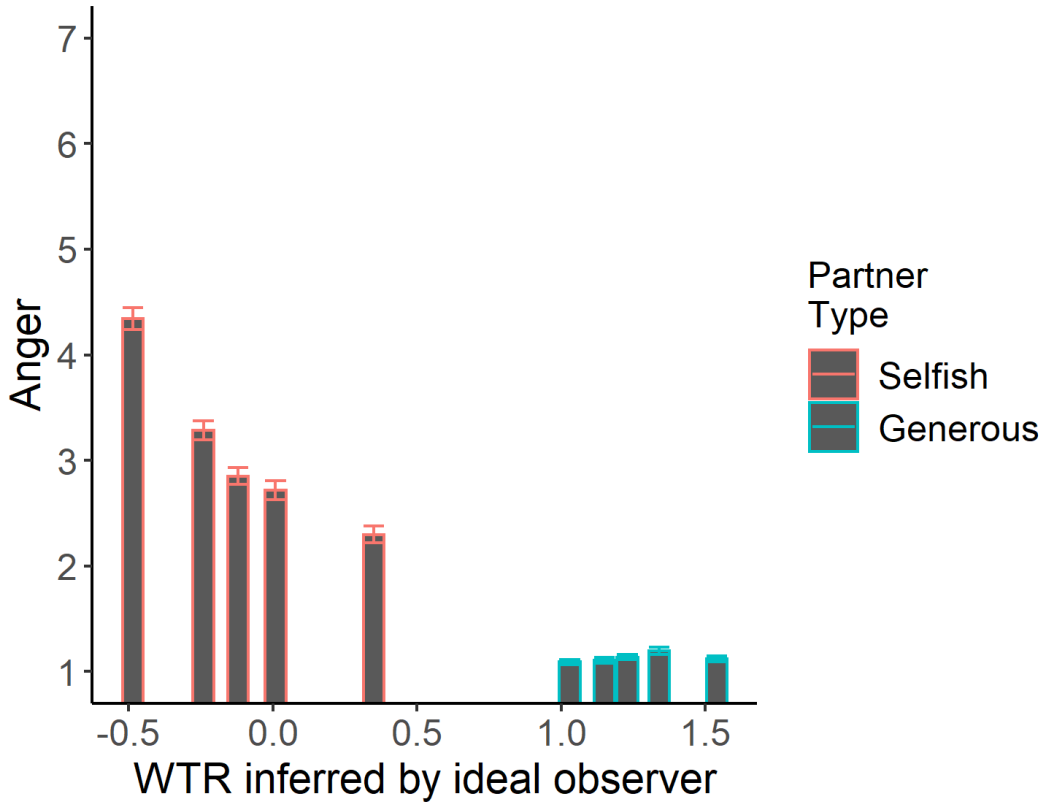


Figure 2.2: Participants' mean anger for each partner as a function of the WTR inferred by the ideal observer for that partner, in Study 2. Error bars represent standard errors of the mean.

the WTR of a stranger) Therefore, the ideal observer calibrated with the prior derived in study 2 might be a more appropriate model of their behavior than the ideal observer calibrated with the first prior I used. Here I reanalyze the data from study 1 with the ideal observer calibrated as in study 2.

2.5.1 Results

Do human predictions match ideal observer predictions?

Yes. The item-level correlation between the average human prediction for a given trial and the model prediction for that trial was r(48)= .978, p <.001. Figure 2.4 shows that both human and model predictions are regulated by the same factors: partners for

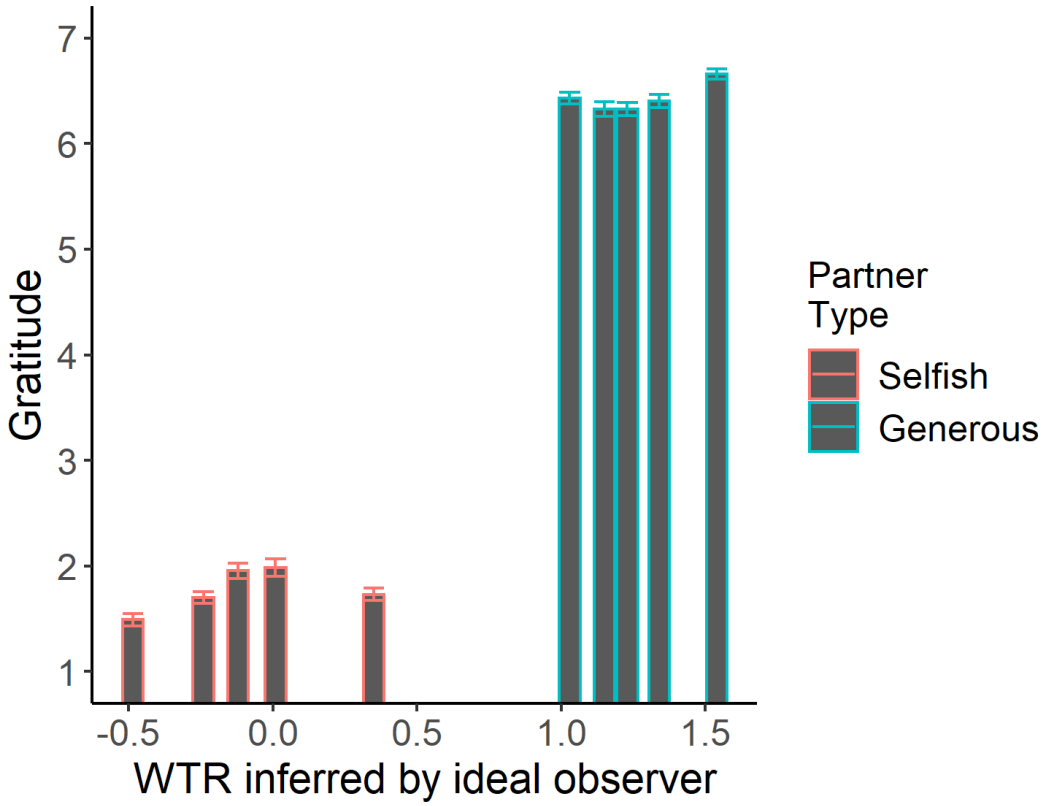


Figure 2.3: Participants' mean gratitude for each partner as a function of the WTR inferred by the ideal observer for that partner, in Study 2. Error bars represent standard errors of the mean.

whom the ideal observer inferred a high WTR elicit more optimistic predictions, and trials for whom the cost of giving was high elicit less optimistic prediction.

Instead of inferences, can human predictions be explained as the result of simple heuristics?

Controlling for material payoffs, the WTR inferred by the ideal observer is positively associated with human predictions, b = .33, p < .001 (linear mixed model with random slopes and random intercepts, material payoffs and inferred WTR as fixed effects, and participant as a random effect). This suggests that inferences about social valuation did play a role in people's predictions.

Does the WTR inferred by the ideal observer predict anger and gratitude?

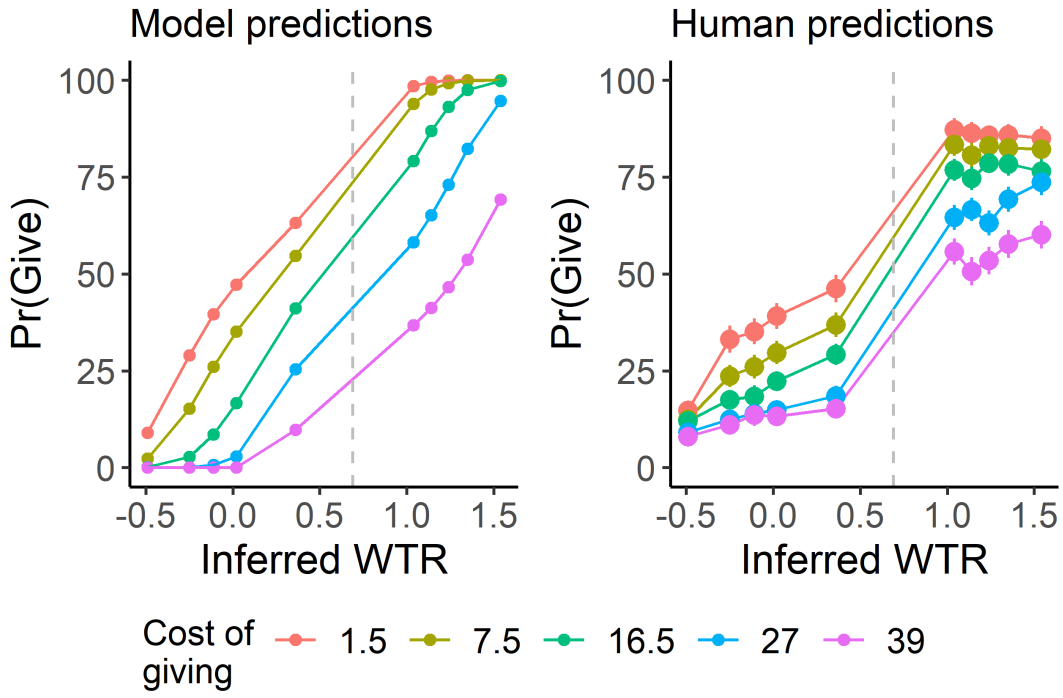


Figure 2.4: Predictions made by the ideal observer (left) and average predictions made by human participants (right), in Study 1. Each dot represents one trial. In both panels, the x-axis represents the WTR that the ideal observer inferred the partner to have toward the participant. "Cost of giving" is the potential payoff (in USD) for the dictator in that trial. Error bars represent standard errors of the mean. Within each panel, selfish partners are at the left of the dashed line, while generous partners are at the right of the dashed line.

Yes for Anger, no for Gratitude. The WTR inferred by the ideal observer was a negative predictor of Anger, b = -.53, p < .001, and a positive predictor of Gratitude, b = .82, p <.001 (linear mixed models with inferred WTR as fixed effect, random slopes and random intercepts, and participant as a random effect).

However, when controlling for material payoffs, inferred WTR was no longer a significant predictor of Gratitude, b = .05, p = .27, though it remained a significant predictor of Anger, b = -.44, p <.001. In other words, variation in gratitude ratings was entirely driven by whether the partner had allocated money to themselves or to the participant. By contrast, participants' anger discriminated even among selfish partners: they were

angrier toward those selfish partners who elicited lower WTR inferences in the ideal observer (see figure 2.5).

Note that in both study 1 and study 2, Gratitude ratings for generous partners were near ceiling, with more than 50% of ratings being on the maximum point on the scale (7 on a 1-7 likert scale) – this may have limited our ability to detect any effect of inferred WTR on Gratitude. Future studies could address this limitation in the design, for example by having participants play with partners who deliver smaller benefits.

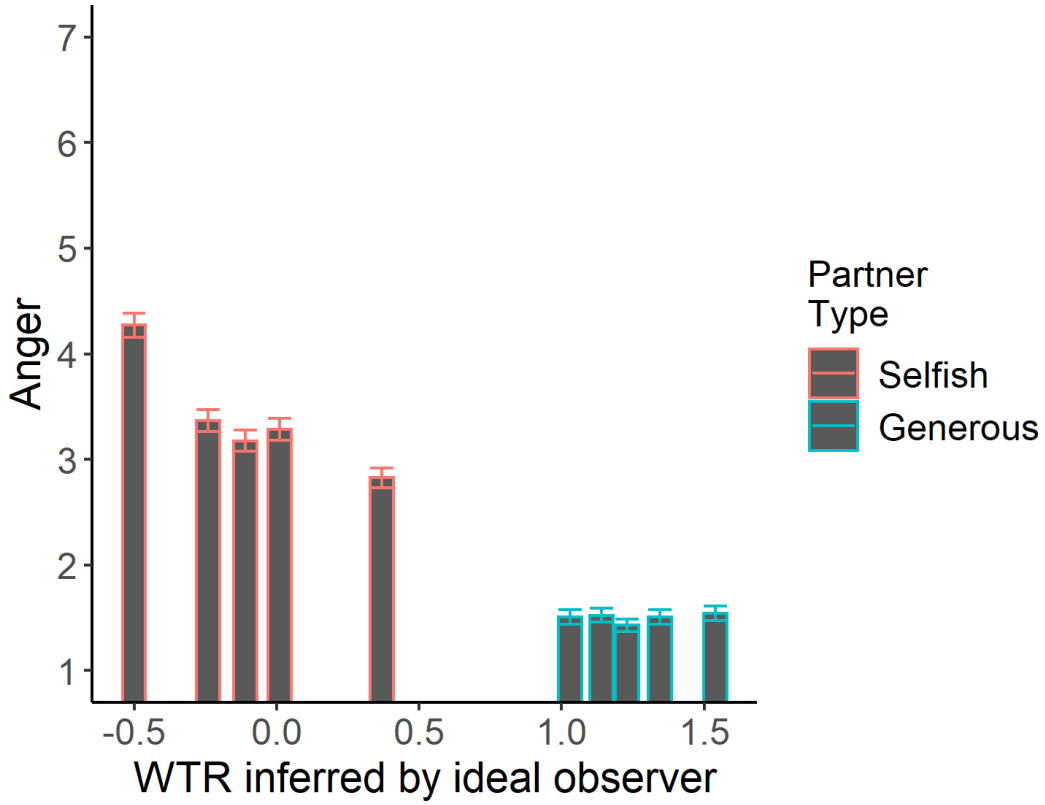


Figure 2.5: Participants' mean anger for each partner as a function of the WTR inferred by the ideal observer for that partner, in Study 1. Error bars represent standard errors of the mean.

2.6 Individual-level analyses

Above, I reported results of study 1 and 2 with item-level correlations, which collapse across participants, and linear mixed models, which partially pool data across participants. Because each participant made 50 predictions and made 20 emotion ratings, it is also possible to treat each participant as its own statistical universe, and perform analyses at the individual level. Doing so provides a robustness check, ensuring that results reported in the main text are not an artifact of averaging, and also gives a sense of the variability between participants. Here I use boxplots to report the results of 650 statistical models (e.g. multiple regressions, correlation tests), each performed on data from one participant. I do so for study 1 and 2, using the ideal observer calibrated with the prior derived in study 2.

2.6.1 Correlation between model predictions and participant predictions

For each participant, I computed the correlation, across trials, between the participant's predictions and the ideal observer predictions. Figure 2.6 reports the distribution of these correlation coefficients, showing that for most participants, there was a close fit between model and participant predictions, especially in study 2 (in-lab sample).

2.6.2 Association between Inferred WTR and participant predictions

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

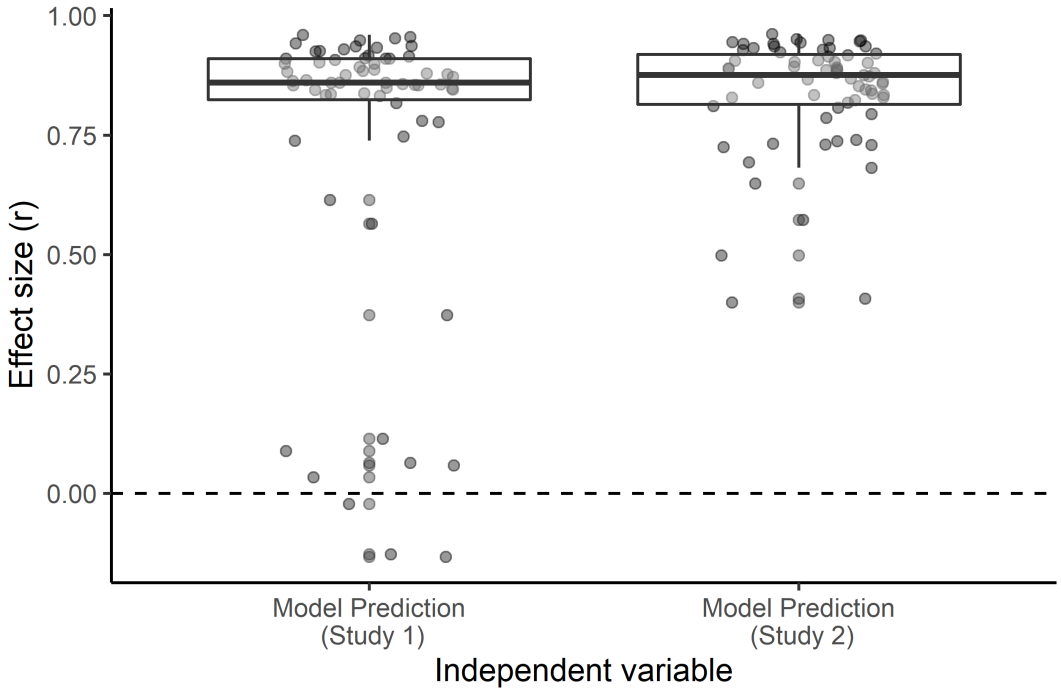


Figure 2.6: Pearson's correlation coefficients (r) for the association between model predictions and participant predictions. Each data point corresponds to one correlation coefficient (i.e. to one participant). Points are jittered along the x-axis for readability.

hold, even controlling for material payoffs.

For each participant, I computed two linear regression models, with ideal-observer-inferred WTR as an IV and participant prediction as a DV. The second model also had material payoffs as an additional IV. For each test, I extracted the standardized regression coefficient for the inferred WTR variable. Figure 2.7 reports the distribution of these coefficients, showing that for most participants, there was a close fit between model and participant predictions, and this association remained, although it was attenuated, when controlling for material payoffs.

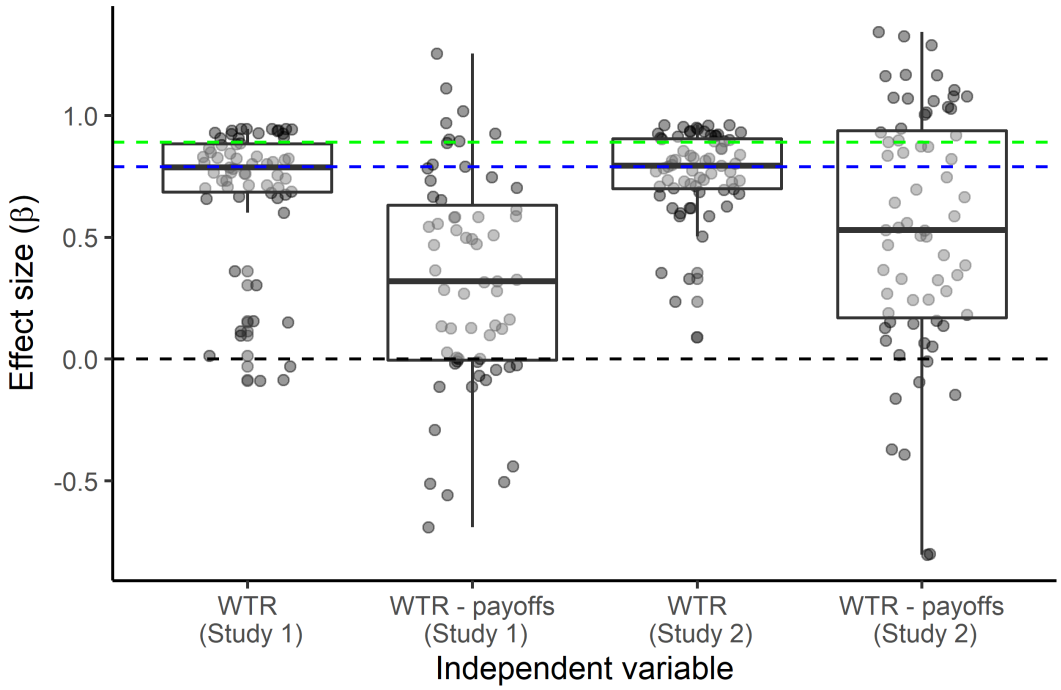


Figure 2.7: Standardized regression coefficients (β) for the association between ideal-observer-inferred WTR and participant predictions. Each data point corresponds to one coefficient (i.e. to one participant). 'WTR': zero-order association between inferred WTR and prediction; 'WTR – payoffs': association between inferred WTR and predictions, controlling for material payoffs. The green dashed line corresponds to the association between ideal-observer-inferred WTR and the model predictions. The blue dashed line corresponds to the same value, controlling for material payoffs. Points are jittered along the x-axis for readability. (The predictions made by the ideal observer model are not perfectly correlated with the WTR that the model infers a partner to have, because the model makes 5 predictions for each partner, and these 5 predictions involve different rounds of the WTT with different payoffs).

2.6.3 Association between Inferred WTRs and emotions

For each participant, I computed four linear regression models, two for Anger and two for Gratitude. For each emotion, the first model had Inferred WTR as an IV and Anger (or Gratitude) as a DV. The second model also had material payoffs as an additional IV. For each test, I extracted the standardized regression coefficient for the inferred WTR variable. Figures 2.8 2.9 report the distribution of these coefficients. For most participants, the WTR inferred by the ideal observer for a partner was a strong predictor of the

participant's Anger and Gratitude toward that partner. For Gratitude, this association considerably weakened when controlling for material payoffs. For Anger, this association remained (on average) unchanged even when controlling for material payoffs.

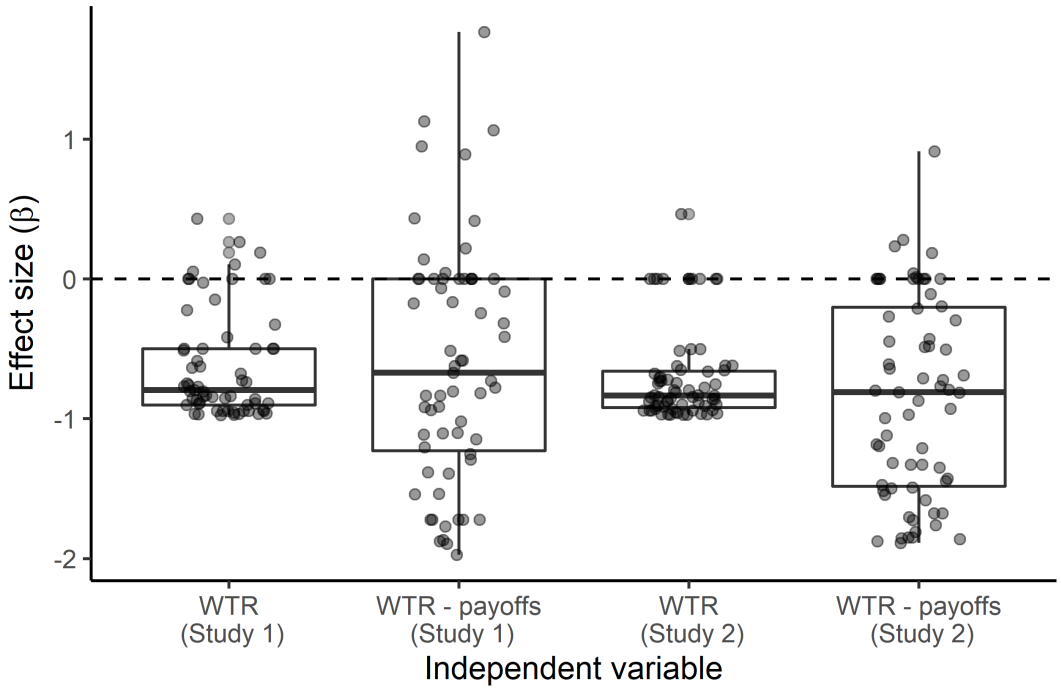


Figure 2.8: Standardized regression coefficients (β) for the association between ideal-observer-inferred WTR and participant Anger. Each data point corresponds to one coefficient (i.e. to one participant). 'WTR': zero-order association between inferred WTR and Anger; 'WTR – payoffs': association between inferred WTR and Anger, controlling for material payoffs. Points are jittered along the x-axis for readability.

2.7 Are individual differences in anger and gratitude explained by different inferences?

According to welfare-inference theories of anger and gratitude, the social valuation inference made by a participant is an input to their emotions. Above we tested this by looking at whether the inferences made by the ideal observer predict participants'

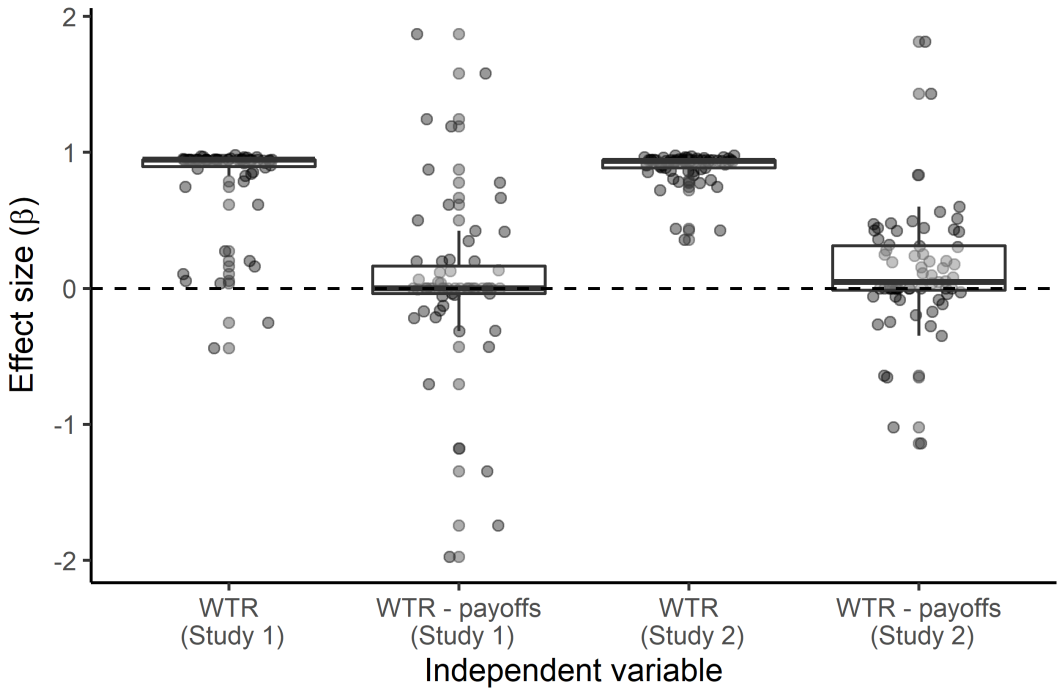


Figure 2.9: Standardized regression coefficients (β) for the association between ideal-observer-inferred WTR and participant Gratitude. Each data point corresponds to one coefficient (i.e. to one participant). 'WTR': zero-order association between inferred WTR and Gratitude; 'WTR – payoffs': association between inferred WTR and Gratitude, controlling for material payoffs. Points are jittered along the x-axis for readability.

emotion ratings. But we can also use participants' predictions as an indirect measure of their inferences: presumably people who make more optimistic inferences about a target inferred that this target had a higher WTR. Therefore it is possible to test the welfare-inference theories of anger and gratitude by looking at whether individual differences in emotion ratings are predicted by individual differences in inferences (as indexed by the predictions a participant made about the target's decisions).

In sum, participant predictions should be correlated with their emotion ratings, even when holding the stimulus constant. That is, for participants observing the same partner, participants reporting higher anger (or lower gratitude) would subsequently predict a lower likelihood that the partner will allocate the money to the participant in observation

trials.

2.7.1 Results

I ran linear mixed models with random slopes and random intercepts, participant prediction as fixed effect, and partner identity as a random effect[5].

In study 1, participant predictions were negatively associated with Anger, b = -.17, p = .009, and positively associated with Gratitude, b = .23, p = .001.

In study 2, participant predictions were negatively associated with Anger, b = -.09, p = .02, but were not associated with Gratitude, b = .004, p = .84.

2.7.2 Results controlling for engagement with the task

The results above provide some support for the hypothesis (for anger, in both studies, and for gratitude in study 1). However, the finding might be explained by the following confound. Consider two participants: participant A pays little attention to the task and responds randomly, while participant B pays close attention. When evaluating a selfish partner, participant B will make less optimistic predictions, and will also report higher anger, than participant A. Therefore, the presence of many inattentive participants would on its own be enough to cause a correlation between predictions and emotion ratings. The fact that I found a stronger effect in study 1, which used online participants who may be more likely to be inattentive, is consistent with this possibility.

In order to control for this possible confound, I computed, for each participant, the correlation between that participants' predictions and the ideal observer predictions across trials. I used this variable as a proxy for the participants' engagement with the

[5]Note that, for a given participant viewing a given partner, we are interested in the average prediction across the 5 predictions that the participant made about that partner, but in a multilevel statistical framework we don't have to actually compute this average.

task. I now report the same analyses as above, with engagement with the task as an additional covariate.

In study 1, participant predictions were negatively associated with Anger, b = -.08, p = .03, but were not associated with Gratitude, b = .04, p = .12. In study 2, participant predictions were negatively associated with Anger, b = -.07, p = .04, but were not associated with Gratitude, b = .004, p = .84.

That is, controlling for engagement with the task, there was no correlation between a participant's predictions and their ratings of gratitude, in either study 1 or study 2. In both study 1 and 2 I found a correlation between predictions and anger ratings, but these were not very far below the conventional threshold for statistical significance. In sum, the current tests provide some (statistically weak) support for the welfare-inference theory in regards to anger, and no support in regards to gratitude. However, as already noted earlier the low variability in gratitude ratings may have made an effect difficult to detect.

2.8 Are individual differences in anger and gratitude explained by individual differences in surprise?

The design of study 2 make it possible to assess, for each participant for which I was able to infer a prior (N=59), how surprising a given set of decisions must have been to that participant.

In the previous sections, I use a single ideal observer, whose prior is an average of the prior I inferred for each participant. In the current section, I created a different ideal observer model for each participant, by equipping that ideal observer with the prior I inferred for that participant.

For every participant and every partner, I computed the surprise that would be experienced by the participant-specific ideal observer, when observing a pair of decisions made by that partner. This yields a different surprise score for each {participant, partner} pair. Formally, I quantify surprise using an information-theoretic measure, the Kullback-Leibler divergence, which quantifies the extent to which a new piece of information causes the ideal observer to update its belief about the partner's WTR (see mathematical details in Appendix D). I also computed the WTR which would be inferred from the partner's decisions by an ideal observer with the same prior as the participant.

The recalibrational theories of anger and gratitude hold that anger and gratitude are, in part, a function of the expectations we have about our partners' WTR: for instance, it predicts that anger should be elicited by cues that a partner's WTR is lower than we would expect (Sell et al., 2017). We can test this prediction by testing whether, in the current sample, individual differences in surprise predict individual differences in anger and gratitude. I do so in an analysis that holds constant the stimuli presented to the participant. In essence, I ask: for different participants looking at the exact same two decisions, are participants who are more surprised by the decisions also angrier (for selfish decisions) and more grateful (for generous decisions)? Because I estimate a participant's surprise as the surprise that would experienced by an ideal observer with the same prior as the participant, this measure does not rely on a direct self-report from the participant, so the tests are unlikely to be contaminated by demand characteristics. Note that the following analyses are exploratory, in the sense that they were not pre-registered before data collection.

I ran two linear mixed models, with random slopes and random intercepts, surprise as fixed effect, partner identity as random effect, and emotion rating as outcome variable. When analyzing reactions to selfish decisions, I found that more surprised participants reported more anger, $b = .40$, $p < .001$. Similarly, when analyzing reactions to generous

decisions[6], more surprised participants reported more gratitude, b = .08, p = .006.

What explains the effect of surprise?

The effect just found can be decomposed in two components. First, by the information-theoretic definition of surprise I use here, agents who are more surprised by a set of decisions are those that updated their beliefs more as a result of seeing these decisions. Thus, very surprised participants may have computed lower WTR estimates (when watching selfish decisions) and higher WTR estimates (when watching generous decisions) than unsurprised participants. These different WTR estimates may in turn have resulted in differences in emotion ratings. Second, it may be that the magnitude of surprise itself, independent of the resulting WTR estimates, regulates emotion ratings.

To see whether each of these components had an independent effect on emotion ratings, I computed two linear mixed models, with random slopes and random intercepts, surprise and estimated WTR as fixed effects, partner identity as random effect, and emotion ratings as outcome variable. When analyzing reactions to selfish decisions, Anger was positively associated with surprise, b = .41, p <.001 (see figure 2.8.1), and negatively associated with estimated WTR, b = -.38, p = .002. When analyzing reactions to generous decisions, Gratitude was positively associated with surprise, b = .11, p = .006, and with estimated WTR, b = .23, p = .006. This suggests that individual differences in both the absolute magnitude of the WTRs estimated by the participants, and in the deviation of these WTRs from their priors, independently predict individual differences in their anger and gratitude ratings.

Are these results explained by a low-level confound?

These results are correlational: instead of an effect of surprise, could they reflect un-

[6]For completeness, I also analyzed emotion ratings in the 'paradoxical' corners: anger toward generous targets and gratitude toward selfish targets. Here one expects that variation in ratings is mostly noise (e.g. inattention), and therefore we do not expect surprise to have an effect. Surprise had no effect on Anger (b=.05, p=.16), or on Gratitude (b=-.03, p=.46). Nine percent of the anger ratings toward generous partners, and 37% of gratitude ratings toward selfish partners, were above 1 (on a 1-7 scale)

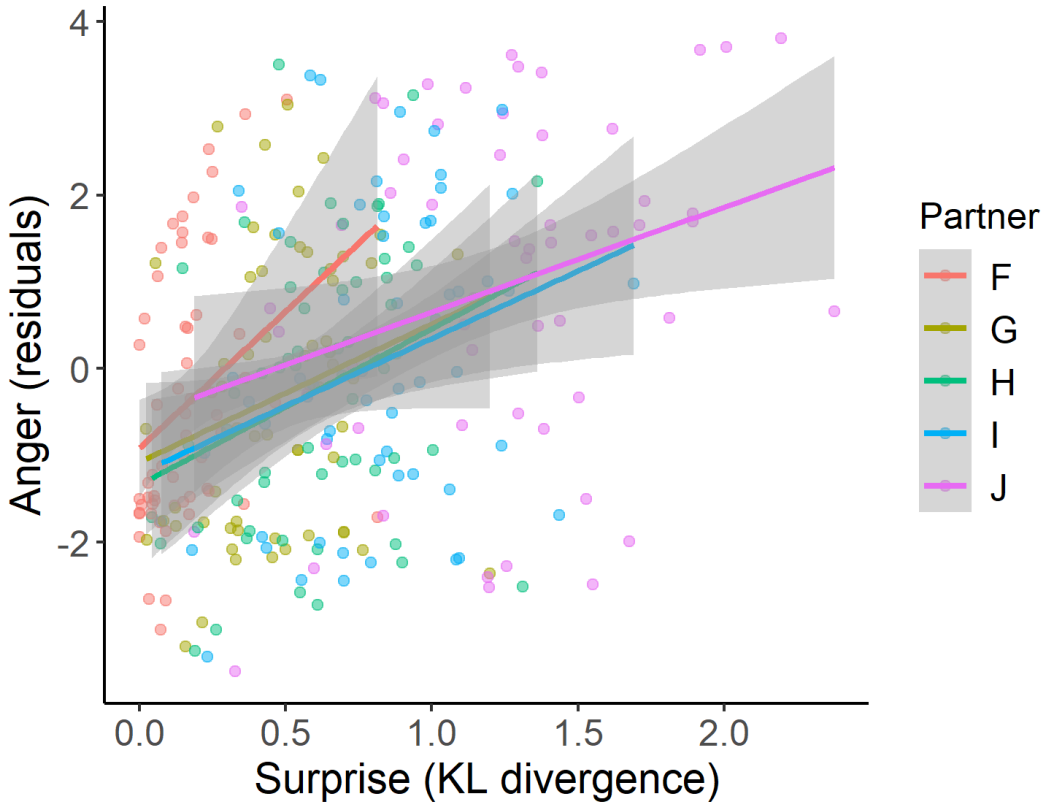


Figure 2.10: Effect of surprise on anger ratings, when estimated WTR is held constant (Study 2). Each regression line corresponds to one partner, and each point is one individual anger rating.

interesting differences between participants? For instance, it may be that the algorithms I use to compute surprise scores assigned the lowest surprise scores to the participants who did not pay attention to the task, and that these are also the participants who report the least anger when observing selfish decisions.

In order to control for this possibility, for each participant I computed an estimate of how much attention the participant paid during the prediction task I used to assess their prior. Theoretically, participants should predict a lower likelihood that their partner would give in WTT trials where the potential benefit for the partner is high. Therefore, if a participant paid attention during the prediction task, their rating of the likelihood that the partner would allocate the money to the participant should be highly nega-

tively correlated with the potential benefit to the partner. Conversely, if the participant responded randomly, we should expect no such correlation.

For each participant, I computed the magnitude of this correlation. I use it in the following analyses as a proxy for the participant's engagement with the task.

I ran linear mixed models with random slopes and intercepts, surprise, estimated WTR, and engagement with the task as fixed effects, partner identity as random effect, and emotion ratings as outcome variable. Analyzing reactions to selfish partners, I found that Anger was positively associated with surprise, beta = .40, p <.001, and negatively associated with estimated WTR, beta = -.16, p = .05. Analyzing reactions to generous partners, Gratitude was positively associated with surprise, beta = .16, p = .01, and with estimated WTR, beta = .18, p = .01.

In sum, there was relatively robust evidence that individual differences in emotion ratings can be explained by two independent factors. First, for a given partner, participants who inferred a lower WTR for this partner expressed higher anger and lower gratitude, regardless of how much they were surprised by that partner's decision. Second, and controlling for the first effect, for a given partner, participants who were more surprised by that partner's decision gave more extreme emotion ratings toward that partner (toward selfish partners: more anger; toward generous partners: more gratitude). Of course, when interpreting these results, it should be kept in mind that I did not directly measure the surprise and the WTR inferences of participants. Instead the current measures of surprise and inferred WTR are computed from the point of view of an ideal observer who would have the same prior as the participant. On the one hand, this is a limitation; on the other hand, this indirect measure is not subject to demand characteristics (e.g. participants using their ratings of surprise as an opportunity to express their anger).

It should be emphasized that these results are exploratory: I did not have these analyses in mind when designing the study. Therefore I conducted a further study to

attempt to replicate the findings. This study was appended as a short task at the end of the experiment reported in the next chapter.

2.9 Study 2B: a replication attempt of the effect of surprise on anger and gratitude

Though the effect of surprise on emotion ratings found earlier is predicted by the social-valuation-inference theories of anger and gratitude, the analysis reported above was not part of the pre-registration. In order to assess the reliability of this exploratory finding, I conducted a conceptual replication.

2.9.1 Participants

Participants were 216 undergraduate students (145 female, mean age = 19.0) recruited from the undergraduate psychology subject pool at a university in California, who completed the study as part of larger experiment (reported in chapter 3). I analyzed data for 134 participants (94 female, mean age = 18.9) who passed an attention check and for whom I was able to compute a prior distribution of WTR.

2.9.2 Procedure

Participants first were asked to make predictions about the WTT decisions of dictators for whom they had no information about past behavior. I used these data to infer the participants' priors. This was the same prior extraction task as described in study 2. In a later phase of the study, participants played the WTT as recipients with 10 computer-generated interaction partners (as in the previous studies, participants were aware that the partners were computer-generated). They observed each partner make one decision

in the WTT, and were asked to rate how angry and how grateful they would feel toward that partner, using two 1-7 likert scales. 5 partners made a selfish decision, and 5 partners made a generous decision ($\pi_{partner}$ was drawn without replacement from {\$5, \$15, \$25, \$35, \$45}, and $\pi_{participant}$ was always \$30). Partners were presented in random order, on a separate page each.

For each decision that a participant saw, I computed the WTR that an ideal observer with the same prior as the participant would infer from this decision, as well as the surprise experienced by that ideal observer.

2.10 Results

I conducted two linear mixed models, with random slopes and random intercepts, surprise and estimated WTR as fixed effects, partner identity as random effect, and emotion ratings as outcome variable. When analyzing reactions to selfish decisions, Anger was positively associated with surprise, b = .44, p <.001 (see Figure 2.11), and negatively associated with estimated WTR, b = -.21, p = .01. The effect of surprise on Anger was robust to controlling for engagement in the task, b = .39, p <.001, but the effect of estimated WTR was not, b = -.08, p = .23.

By contrast, when analyzing reactions to generous decisions, Gratitude was not associated with surprise, b= .02, p = .30, or with estimated WTR, b = .09, p = .08.

As a sanity check, I also performed similar tests for the "paradoxical corners": anger ratings toward generous decisions, and gratitude ratings toward selfish decisions (recall that for each partner, I asked participants to rate both the gratitude and the anger they would feel toward that partner, regardless of whether the partner was "selfish" or "generous"). Surprise had no effect on Anger toward generous partners, b = .00, p=.37. By contrast, as shown in Figure 2.12, Surprise was positively associated with Gratitude

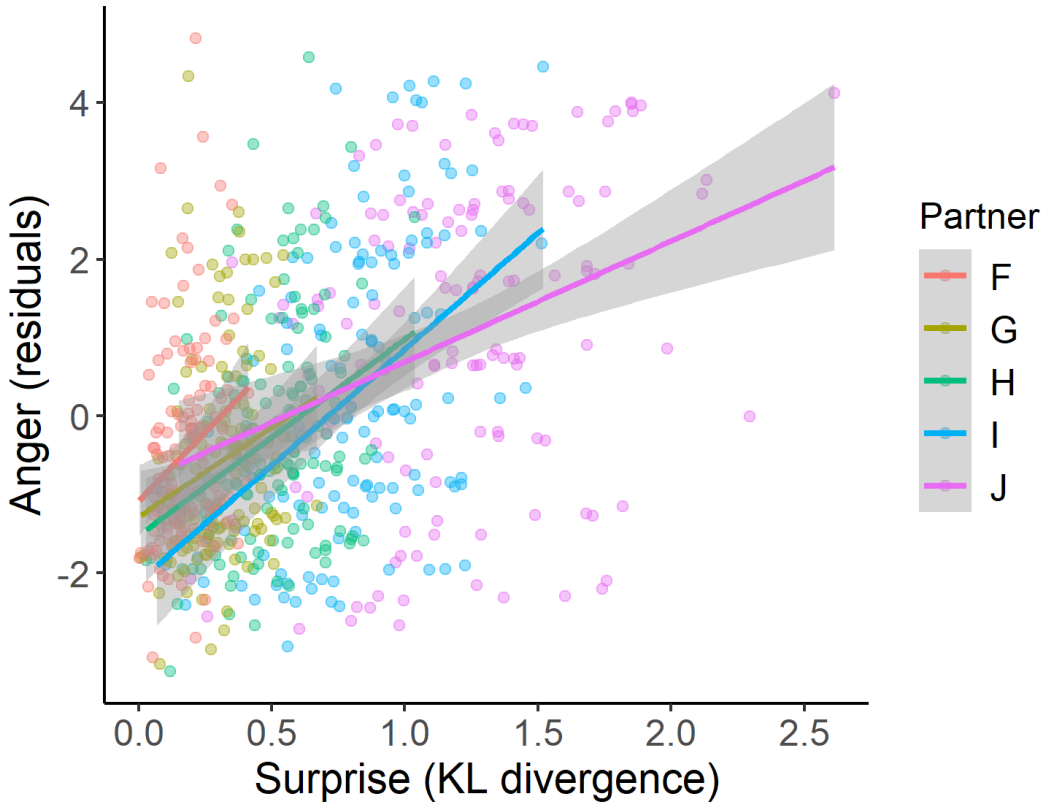


Figure 2.11: Effect of surprise on anger ratings, when estimated WTR is held constant (Study 2B). Each regression line corresponds to one partner, and each point is one individual anger rating.

toward selfish partners, b = .40, p < .001 (linear mixed models with partner identity as random effect). In the same model with estimated WTR and engagement with the task as covariates, Surprise was still a positive predictor of Gratitude, b=.53, p < .001.

The latter result casts some doubt on the validity of the approach used here to quantify surprise. Theoretically, one would expect positive gratitude ratings toward selfish partners to reflect either noise, or participants using their answer to the gratitude question to communicate that they are not very angry. If the latter, then one would expect surprise to be a *negative* predictor of gratitude (since by hypothesis very surprised participants would feel more anger toward a selfish partner). This suggests that my measure of surprise may be picking up unrelated noise, such as variation in participant's

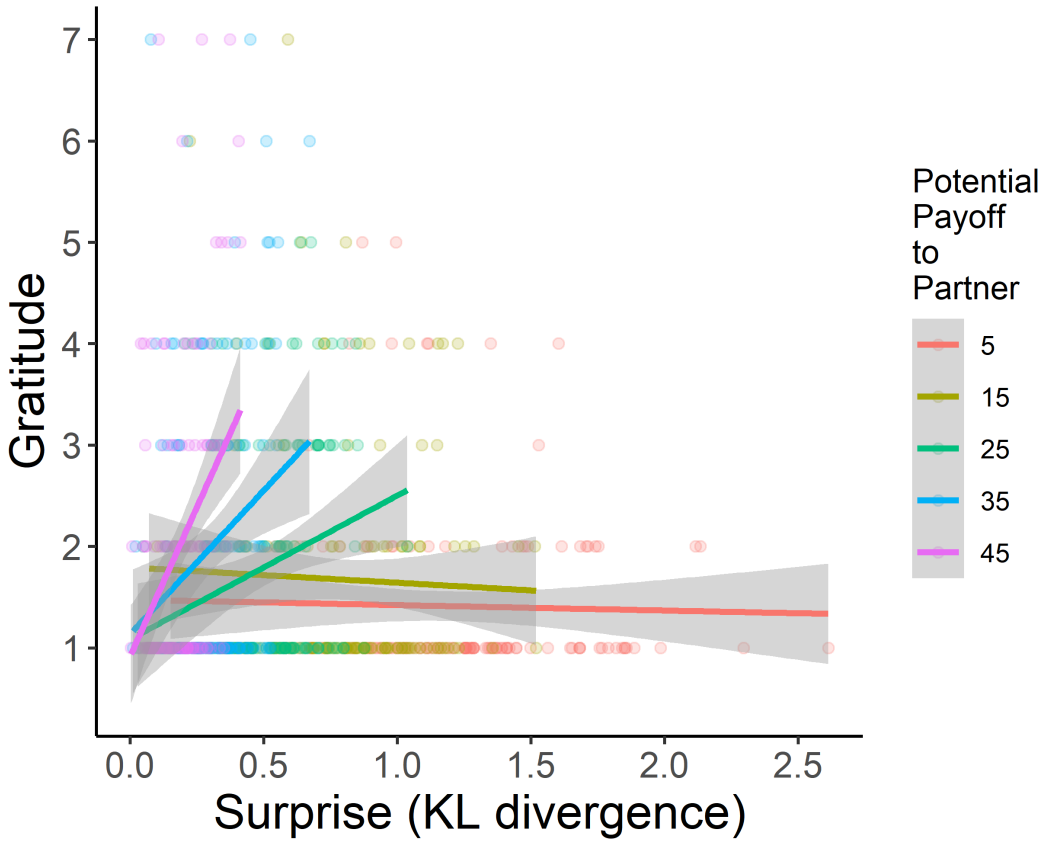


Figure 2.12: Effect of surprise on "paradoxical" gratitude ratings, i.e. gratitude ratings toward selfish partners (study 2B). Each regression line corresponds to one partner, and each point is one individual gratitude rating.

engagement with the task. Although I attempted to control for engagement with the task (by using the correlation between a participants' prediction and $\pi_{dictator}$ in the prior extraction phase), the fact that surprise still predicts gratitude toward selfish partners even controlling for that measure suggests it might be imperfect.

In sum, data from study 2 and study 2B suggest that surprise might intensify emotion ratings, although the validity of the current measure of surprise is uncertain.

2.11 Discussion

When predicting the behavior of others in a task involving welfare-tradeoffs, participants made predictions that closely tracked the predictions made by a Bayesian ideal observer for this task.

For each person they had to evaluate, participants could only observe two decisions that this person made. Often these decisions did not contain enough information to allow straightforward predictions about how the person would behave in other contexts. Therefore, participants had to solve a difficult problem of statistical inference under uncertainty. The close fit between their behavior and that of the ideal observer is surprising from the perspective of the large body of work documenting that humans systematically deviate from normative statistical reasoning in many contexts (Kahneman, Slovic & Tversky, 1982; Marcus, 2008). On the other hand, these findings provide additional evidence that, in ecologically valid contexts, human statistical inference can approximate Bayesian standards (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995; Knills & Richard, 1996; Griffiths & Tenenbaum, 2006; Weiss, Simoncelli & Adelson, 2002).

Note that a domain-general ability to draw sound statistical inferences would not be enough, on its own, to generate the kinds of judgments that participants made. The ideal observer model also relies on a set of domain-specific assumptions about the way people typically make welfare trade-offs. The causal model used by the ideal observer assumes that agents maximize a utility function containing parameters that regulate the relative weight that the agent assigns to the welfare of the participant relative to its own. The tight fit between model and human behavior suggests that participants had access to a similar kind of domain-specific knowledge. Thus, humans probably represent the minds of other agents as containing such welfare trade-off parameters (Tooby et al., 2008, Sell et al., 2017).

Analysis of participants' anger ratings provide convergent evidence that people infer the value of the welfare trade-off parameters in the minds of others. Partners who made decisions that imply a lower valuation of the participant (as assessed by the ideal observer) elicited more anger. Importantly, this was true even when holding constant the total opportunity costs that each partner inflicted on the participants, and the total gains that each partner obtained at the expense of the participants. I found only weak evidence for an association between welfare valuation inference and gratitude, although this might be due to a strong ceiling effect for gratitude ratings – the extent to which gratitude depends on valuation inferences remains an important area for future research.

I also looked at whether social-valuation-inference theories of emotions could explain individual differences in anger and gratitude. There was (statistically weak) evidence that among participants reacting to the same partner, participants who made less optimistic predictions toward that partner (an index of low-WTR inference) were angrier at that partner. I found no evidence for such an effect for gratitude ratings (although again this may be due to a ceiling effect).

Finally, in exploratory analyses, I computed an individualized estimate for the prior of each participant in study 2, thus creating a different ideal observer for each participant, calibrated to the (estimated) prior of that participant. This allowed me to compute indirect measures of the surprise felt by a participant in response to a partner's decision, and the WTR that the participant estimates this partner to have. Individual differences in these measures explained some of the individual differences in emotion ratings, in the predicted direction: surprise exacerbated the intensity of emotions, while participants who estimated low WTRs for a partner were angrier and less grateful. In a subsequent experiment, I successfully replicated the association between surprise and anger ratings, but not the other results. That experiment also revealed a positive association between surprise and gratitude ratings toward selfish partners, which suggests low-level confounds

may be explaining the effect of surprise on emotion ratings.

Therefore, the current data on surprise should be interpreted as suggestive evidence, calling for further research. Convergent evidence for the role of surprise (in the information-theoretic sense used here) would support a strong version of the social-valuation-inference thesis: emotions are not a simple readout of the WTR estimate that results from observing a decision; they also reflect the extent of the shift in that estimate that happens during inference.

In the two studies reported in this chapter, participants were passive observers. In the next chapter I test the active component of learning: do people know how to look for the evidence that is most information-rich?

Chapter 3

Are humans rationally curious about social valuation?

Suppose 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. From the point of view of the net profit model, there is no reason why this question should raise your curiosity: whether she 'recycled' the ticket does not affect its objective value. However, this piece of information would tell you something about how much she values you.

The experiment reported in the current chapter investigates whether people are curious about the sort of evidence that reveals how much others value their welfare. It uses a quantitative approach, that expands on the ideal observer method used in the previous chapter.

3.1 Theoretical framework

Humans are active learners: they do not just passively observe the world, but they explore it, selectively seeking out the information that has the most potential to expand their knowledge. As young as 11 months, infants who watch toys that violate their core knowledge intuitions will spontaneously manipulate them in ways that would allow them to learn more about the violation (Stahl & Feigenson, 2015). Active learning typically gives more information than passive learning to the learner: people who had to learn the properties of objects in a simulated 'microworld' were quite successful if they were able to directly interact with the object; 'yoked' participants who passively saw the video clips generated by the active learners did not learn as much (Bramley, Gerstenberg, Tenenbaum & Gureckis, 2018). The mere act of moving one's eyes is an act of active learning, directing our gaze to the part of the visual world that contains the most relevant information (Najemnik & Geisler, 2005).

There are at least two reasons why active learning is better than passive learning. The first is familiar to any freshman taking Intro to Research Methods: because they directly manipulate the variable of interest, experiments allow one to establish the direction of causality with greater confidence. The second reason is that the world does not go out of its way to make us learn: among all the possible observations we could make, only a few of them contain fitness-relevant information. To be most efficient at information-gathering, people should try to put themselves in a position where they get the information most relevant to their epistemic goals. Here I study this second aspect of active learning: how people gather the most information-rich evidence.

3.1.1 In general, are humans good at selecting information-rich data?

The question of the rationality of human data acquisition has a relatively similar answer to the question of the rationality of human statistical inference. Namely, while humans are far from always being able to select the most informative data, they show a very good fit to normative models when the problem is presented in a format that fits the relevant cognitive mechanisms.

Early studies (e.g. Wason, 1966; 1968) highlighted systematic deviations from normative principles of data selection. A flagship study in the 'irrationality' of human data acquisition is Wason's experiments with the selection task that bears his name (Wason, 1966). In the selection task, participants are presented with four cards (e.g. [A], [C], [4], [7]), and are asked to test a rule (for instance "each card with a vowel has an even number on the other side") by turning over as few cards as possible. The correct answer is to turn over the 'A' and '7' cards, but most participants turn over the 'A' and '4' cards. This seems to show that humans fall prey to a 'confirmation bias': instead of choosing the '7' card, which could falsify the rule (if we find a vowel on the other side), people choose the '4' card, which would provide evidence consistent with the rule if there is a vowel on the other side, but otherwise provides no information directly relevant to the goal (even if we find a consonant on the other side, that card would not violate the rule). Hence human data selection is irrational.

The generality of that conclusion has been challenged from several angles. Here is an argument slightly modified from Oaksford & Chater (1994). The mind probably has not been under any strong selection pressure for solving artificial logic puzzles, so it is worth considering the problem that the mind is actually trying to solve when asked to test a rule. For concreteness, let us consider the following scenario: you have been asked to

evaluate the rule "when someone eats tripe, that person gets sick". There are four cards in front of you, representing four people: the first person ate tripe, the second didn't, the third is sick, the fourth one isn't. The problem that the mind is trying to solve here is likely to differ from the original intent behind the logic puzzle in two important ways.

First, the falsificationist answer (turning over [Ate tripe] and [Isn't sick]) gives you all the information you need on the assumption that you want to check whether the rule is valid within the four people considered here, considered as their own universe. But a well-designed mind would want to check the general truth of a rule, here, whether in general people who eat tripe tend to get sick. If you find that the person who ate tripe got sick, and the person who didn't get sick didn't eat tripe, you have checked that the rule holds for the four people on the cards, but it is still possible that other people violate the rule[1]. Turning over the two other cards might reveal useful information about whether the rule applies in general.

Second, the rule that the participant has to check is a deterministic rule, which could be written, in the formalism of probability theory, $P(is\,sick|ate\,tripe) = 1$. In the real world, causal relationships are rarely deterministic, and non-deterministic rules can matter a lot. If $P(is\,sick|ate\,tripe)$ was .9 instead of 1, you would still want to know it before you order food. Therefore it is likely that the mind is designed to infer the strength of a relationship, instead of simply checking whether a rule always holds.

Assuming that your mind is designed to infer the strength of not-necessarily-deterministic relationships, that can generalize out-of-sample, which cards should you turn over? It depends on the probability of events. Here, it is reasonable to assume that both events (eating tripe and getting sick) are relatively rare.

Is it useful, then, to turn over the card [isn't sick]? Well, since most people don't eat

[1]More generally: even if these four people were the only humans left on Earth, you may still be interested in the causal effect of eating tripe, beyond what actually happened to the people in your sample, for instance to predict what might happen to them in the future (see Quillien, 2015).

tripe, finding that the non-sick person didn't eat tripe is not very surprising. Finding that the person ate tripe would be highly informative (since it violates the rule), but it is a relatively unlikely outcome of the observation, so overall the expected information gain of turning that card over is relatively low.

By contrast, it is useful to turn over the card that says [Is sick]. If you find [Ate tripe], this would constitute strong evidence in favor of the rule. Here is why. Eating tripe is a rare event, and getting sick is a rare event. Assuming statistical independence between these events (or a very weak relationship), people to which both events happen should be very rare. Therefore finding a sick person who ate tripe constitutes strong evidence for a relationship between eating tripe and getting sick (see Oaksford & Chater, 1994, for proof)[2]. The [is sick] card contains more potential information than the [is not sick] card, despite the fact that the falsificationist considers only the second to be of any use!

In sum, assuming that the mind is designed to infer non-deterministic rules that generalize, and that they assume that the relevant events (or properties) are rare[3], the pattern of behavior of participants in a Wason selection task is optimal.

In other versions of the selection task, the mind may be trying to infer something other than the strength of a general causal relationship. Cosmides and her colleagues have argued that the mind hosts specialized mechanisms to reason about specific domains such as precautions and social exchange, and that these mechanisms are activated by some versions of the selection task.

Evolutionary game-theoretic considerations suggest that for social exchange (i.e. reci-

[2]Oaksford & Chater have a slightly different model than the one I use here. For mathematical convenience, they assume that people consider only two hypotheses, either $P(is\,sick|ate\,tripe) = 1$ or $P(is\,sick|ate\,tripe) = P(sick)$; that is, either sickness and tripe are completely statistically dependent or they are completely independent. It is a natural extension to assume instead that any degree of statistical dependence between these two extremes is a plausible hypothesis from the participants' point of view.

[3]Oaksford & Chater (1994) provide some reasons to think that a tendency to make the rarity assumption might be ecologically rational.

procity) to be an evolutionarily stable strategy, organisms need to have mechanisms to make their cooperation contingent on the cooperation of their partner. That is, they need a way to detect cheaters. Since social exchange is observed in humans, the human mind must host cognitive mechanisms that are able to detect cheaters (Cosmides, 1989).

A task analysis of cheater detection (Cosmides, 1985) makes predictions about the design of such cognitive mechanisms. One expects these mechanisms to be designed to look for data that make it possible to identify cheaters, defined as individuals who intentionally violate a social contract.

This predicts that human performance in the Wason selection task should be high if the task is framed in a way that activates cheater detection algorithms. For instance, if your goal is to check for violations of the rule "if a teenager borrows their parents' car, they must fill up the tank afterwards", by turning over the cards [Borrowed car], [Did not borrow car], [Filled up tank], [Did not fill up tank], the only two cards that allow you to detect a cheater are [Borrowed car] and [Did not fill up tank]. People in fact do select these two cards at highest frequency (Cosmides & Tooby, 2005). Later experiments showed that framing the selection task as a social contract boosts performance mostly when data selection would allow one to detect cheaters: in a version of the task where potential rule violations are said to be unintentional mistakes, people show poor performance (Cosmides, Barrett & Tooby, 2010).

In sum, human 'irrationality' in the Wason selection task is probably not explained by the hypothesis that humans lack cognitive mechanisms for optimal data selection. Instead, the standard version of the task elicits poor performance because it triggers mechanisms designed to solve a different adaptive problem (e.g. inferring the strength of a general causal relationship) than the logical problem originally intended. When the task is framed in such a way that the 'logically correct' answer also matches the inference rules of a specialized inference system, humans succeed at the task.

Instead of comparing human data selection to the standards of formal logic, as Wason did, later work used the frameworks of information and probability theory (see section 3.2), since these are most relevant to the inference problems that organisms typically have to solve. In explicit information-gathering tasks (for instance, where participants have to play a 20-question game, or play the game of 'Battleship', or learn to form categories), researchers have found that normative models of data selection are a generally good description of human behavior, although in some contexts people are not perfectly optimal and use heuristics that only approximate the computational-level solution (Liefgreen, Pilditch & Lagnado, 2020).

Finally, at the level of perception, human eye movements in various tasks have been found to be very close to those made by an ideal search model that maximizes information intake (Najemnik & Geisler, 2005; Nelson & Cottrell, 2007; Peterson & Eckstein, 2012; 2013).

Do we then expect human data selection to be optimal when it comes to gathering information about social valuation? The answer to that question is similar to its analog for passive statistical inference. Given the hypothesis that humans have cognitive systems that make inferences about the welfare-tradeoff parameters of others, it seems likely that evidence that is potentially relevant to social valuation inference will be recognized as such by these systems. These systems should be able to compute the expected information value of a given query, and use these computations to guide data selection.

3.2 Formal models of optimal data selection

Given a causal model of the world, and a set of beliefs about the probable values of the variables in this causal model, it is possible to define the expected informational

value of a query (i.e. an eye movement, a question, an experiment, etc)[4]. To give a simple example using the light bulb example from Chapter 1, suppose you would like to know whether the room is dark, and you are allowed to ask either whether the switch is off or whether the lightbulb is dead. Asking whether the switch is off is the question with the most information value here, since it is the question that will most reduce your uncertainty about whether the room is dark (if you learn that the switch is on, then you know that there is a .99 chance that the room is lit; if you learn that the switch is off, you know with certainty that the room is dark).

Formally, the expected information value of a query (i.e. the act of looking somewhere, asking a question, performing an experiment, etc) can be computed as:

$$EIV = \sum_i Pr(d_i)U(d_i)$$

Where $Pr(d_i)$ is your estimate of the probability that the query will yield datum d_i, and $U(d_i)$ is the informational value of observing d_i. In other words, the expected information value of a query is simply a weighted mean of the information value of its possible outcomes, weighted by the probability of each possible outcome.

For instance, the query "is the switch on?" has two potential outcomes: either I learn that the switch is on, or I learn that the switch is off. My expected information value from this query is:

$$Pr(switch\,is\,on)U(switch\,is\,on) + Pr(switch\,is\,off)U(switch\,is\,off)$$

Where $Pr(switch\,is\,on)$ is my prior estimate of the probability that my question will reveal that the switch is on, and $U(switch\,is\,on)$ is a measure of how useful it is for me

[4]Some researchers refer to this subfield of statistics as Optimal Experimental Design. The name is slightly unfortunate since the theory is not restricted to experiments, and also applies to the informativeness of observational data.

to learn that the switch is on, given my goal of finding out whether the room is dark.

This definition of the expected information value of a query raises a key question: How should $U(d)$, the information value of observing a given datum, be measured?

3.2.1 Proposed measures of the information value of a datum ($U(d)$)

Several different measures of the information value of an observation outcome exist (see Nelson et al., 2010, for review).

Two early measures of information value, Bayesian diagnosticity and log-diagnosticity, have been shown to be a poor index of information value, both on theoretical grounds and in terms of accounting for human information search (Nelson, 2005). Other measures, such as Probability Gain (Baron, 1985), provide good descriptions of human behavior in simple information-gathering problems (Nelson et al., 2010), but in some settings they have counter-intuitive properties that are undesirable for a normative theory. Probability Gain measures by how much the new data increase your estimate of the probability that the hypothesis that you favor (i.e. the one to which you assign highest probability) is actually correct. Formally it is defined as:

$$PG = max_i(P(h_i|d)) – max_i(P(h_i))$$

For instance, let us say you initially think that the room is dark with probability .505. If you are asked whether the room is dark you should answer 'Yes', and you expect that you would be correct 50.5% of the time. After learning that the switch is on, you now believe that there is only a 1% chance that the room is dark (i.e. the 1% chance that the lightbulb is dead): you should now answer 'No' when asked if the room is dark, and you expect that you would be correct 99% of the time. Therefore your probability

gain from the data—learning that the light is on— is .99 - .505 = .485. In other words, your estimate of the likelihood that you will be correct when answering a question about whether the room is dark increased by .485 when learning that the switch is on.

The problem with PG is that it assigns an information value of 0 to evidence that intuitively is useful (Liefgreen, Pilditch & Lagnado, 2020). To see why, imagine a slightly different lightbulb scenario, where the prior probability that the switch is off is 2/3, and you have a very cheap lightbulb whose prior probability of being dead is 25%. In this scenario, simple computations (left as an exercise to the reader) show that you are initially 75% confident that the room is dark, and learning that the switch is on makes you 25% confident that the room is dark.

Therefore, initially you estimate that you have a 75% chance of getting the correct answer (you will say "Yes" to the question "is the room dark?", and you think you will be correct 75% of the time), and after learning that the switch is on you still think that you have a 75% chance of getting the correct answer (you will say "No" to "is the room dark?" and expect to be correct 100 – 25 = 75% of the time). Therefore your probability gain is .75 - .75 = 0. But intuitively learning the state of the switch was useful information with respect to your goal of learning whether the room is dark. Information Gain (Lindley, 1956), a similar measure, suffers from the same problem.

In the current study I will use a widely-used measure of information value, the Kullback-Leibler divergence (Kullback & Leibler, 1951). The KL divergence measures how much your probability distribution over the space of possible hypotheses shifts in response to the observation. Formally it is defined as:

$$KL = \sum_i P(h_i|d) \log \left(\frac{P(h_i|d)}{P(h_i)} \right)$$

Where $P(h_i)$ is the prior probability you assign to hypothesis h_i, and $P(h_i|d)$ is your

posterior belief in h_i given the observation d.

For instance, in the scenario where learning that the switch is on makes you go from a 75% belief to a 25% belief in "the room is dark", we have, for the hypothesis 'the room is dark':

$$P(room\,is\,dark|switch\,is\,on)\log\left(\frac{P(room\,is\,dark|switch\,is\,on)}{P(room\,is\,dark)}\right)$$
$$= .25\log\left(.25/.75\right) = -.27.$$

And for the hypothesis 'the room is lit':

$$P(room\,is\,lit|switch\,is\,on)\log\left(\frac{P(room\,is\,lit|switch\,is\,on)}{P(room\,is\,lit)}\right)$$
$$= .75\log\left(\frac{.75}{.25}\right) = .82$$

Thus we have KL = .82 +(- .27) = .55: we did learn something by observing that the switch is on. In sum, KL divergence does not suffer from the problem that plagues Probability Gain. Empirically, analysis of the data from several papers on human information search shows that KL divergence may be the normative theory which best accounts for human behavior (Nelson, 2005).

In sum, I will use KL as the measure of the information value of a datum, —that is, $U(d)$. The information value of a datum is, in turn, a building block of the measure of the expected information value (EIV) of a query. As a reminder, the expected information value of a query is a weighted mean of the information value of all its possible outcomes (where each outcome is weighted by its estimated probability).

3.3 An ideal search model of WTR inference

Here I define an ideal search model that computes the expected information value of a query about Alice's decisions in the WTT. The task is the following: if you know that Alice had to play a trial of the WTT, and know the payoff involved in that trial, but not whether Alice chose to Give or Take, how much information would you rationally expect to gain by asking about Alice's decision?

For instance, if Alice had to make a choice between \$30 for herself and \$5 for Bob, asking what Alice did is intuitively not very informative about her WTR toward Bob (she could value Bob highly yet still take the \$30). By contrast, if Alice had to make a choice between \$15 for herself and \$30 for Bob, asking about her decision is intuitively very informative. A priori it seems that she could decide either way, and whatever she decides we will have learned something about how much she values Bob. The ideal search model quantifies exactly how informative a question is, and makes queries as a function of its estimate of the informativeness of the different possible questions it could ask.

The ideal search model is a straightforward extension of the ideal observer defined in Chapter 2. Indeed, the information value of an observation is defined in reference to how it changes our beliefs; this value can only be computed by reference to a procedure for updating one's belief. Therefore, here I define information value by reference to how much a given piece of information changes the ideal observer's belief about Alice's WTR.

Specifically, I define the information value of an observation as the KL divergence of the ideal observer's posterior belief from its prior belief, in response to this observation. Figure 3.1 explains the intuition for what KL divergence is measuring in this context. Here the space of all possible hypotheses is continuous. There are an infinity of possible hypotheses about Alice's WTR towards Bob (it could be .4012, .4013, .4014, etc), and this continuum of hypotheses is plotted on the x-axis. The blue curve is an agent'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 agent's posterior belief after having seen a decision by Alice (here, a decision where she chose to Give). Intuitively, the KL divergence measures how much you have to 'move' the blue curve in order to make it into the green curve.

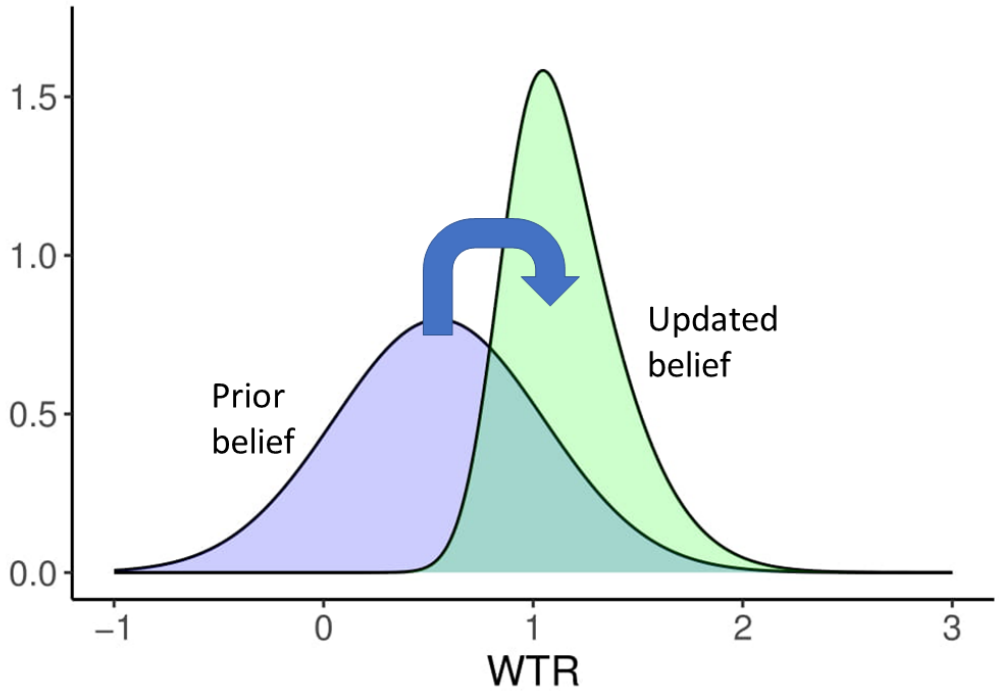


Figure 3.1: Conceptually, the KL divergence quantifies how much one must move the probability distribution corresponding to the observer's prior belief to obtain the belief that has been updated by the observation. The greater the divergence of the updated belief from the prior belief, the higher the information content of the decision. Y-axis: probability density.

Because the hypothesis space is continuous, when formally writing the formula for KL divergence we use an integral sign instead of a summation sign:

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

With KL as our measure of the information value of observing a given datum, that

is, $U(d)$, one can calculate the measure of interest: the expected information value of a given query. The expected information value of a query is simply the weighted mean of the information value of its possible outcomes:

$$EIV = \sum_i Pr(d_i)U(d_i)$$

$$= KL(Take)Pr(Take) + KL(Give)Pr(Give)$$

To give a concrete example, imagine that Alice had to decide between getting \$60 or letting Bob get \$20. If we see her make the generous decision, we gain a lot of information, i.e. Alice is much more generous than we expected. But note that Alice could value Bob very highly yet still take the \$60—the benefit to Alice of the selfish decision (\$60) is three times the benefit to Bob of the generous one. So forgoing \$60 to give Bob \$20 is a priori unlikely. With these payoffs, it is much more plausible that Alice will make the selfish decision, in which case we will not have learnt much. Therefore the expected information value for this trial is low. By contrast, in a trial where Alice has to decide between \$10 for herself and \$20 for Bob, it is not obvious what she will do. She would forgo \$10 to give him \$20 if the weight she puts on Bob's welfare is a bit more than half the weight she puts on her own – which is not too unlikely. Therefore we will have learned something about her WTR no matter what she actually chooses, and this trial has high expected information value.

3.4 Study 3.1

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. Importantly, 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.

3.4.1 Participants

I recruited 216 participants from the undergraduate psychology participant pool at a university in California, 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).

3.4.2 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 to 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 first saw the partner make one decision. The decision made by the partner 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 (on two 1-7 likert scales).

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 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 $\pi_{dictator}$, resulting in fifteen different pairs of trials.

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 tasks that were designed as a replication attempt of the findings about surprise from study 2 chapter 2.

The first task was a prediction task where participants were asked to predict the behavior of other players in the WTT[5]. 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 (see section 2.9 in Chapter 2 for more details). All participants completed the emotion ratings task after the data selection task. The results of this replication attempt are reported in Chapter 2 (Study 2B, section 2.9).

Then participants were asked a few demographic questions and were thanked for their participation.

[5]In hindsight, this prediction task could also have been designed with the goal of calibrating the ideal observer used in the data selection task (just as in study 2 in Chapter 2). However, in its current form the prediction task was not appropriate, because to model participant behavior one needs to know their prior for negative WTRs, but the prediction task used here does not allow one to infer someone's prior for negative WTRs. Therefore for the data selection task I simply calibrated the ideal observer's prior by fitting it to the participants' behavior in the data selection task (see below).

3.4.3 Computational modeling

I compared participants' selections with a stochastic version of the ideal search model. The motivation for this choice is that, 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 gain for observing the outcome of dilemma 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.

In addition to the ideal search model, I tested three alternative computational models of data selection. All models were built on top of the Bayesian 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 is not allowed to observe the one decision that participants observed before the data selection phase of the experiment. Therefore this model makes the exact same choices in the 'Take' and the 'Give' condition.

Following a popular simplification of the philosophy of Karl Popper, many scientists think that information search should ideally consist in a process of 'falsification', but

that people fall prey to a 'confirmation' bias (see Wason, 1966). Accordingly, I tested a 'Falsification' and a 'Confirmation' model of data selection. I note that what counts as 'Falsification' or 'Confirmation' in the context of information-gathering for estimating a continuous value is somewhat ambiguous, so that there could in principle be many ways that one could operationalize such a strategy. Here I choose to test a very simple implementation.

In the 'falsification' model, the agent tries to falsify the hypothesis about the partner's WTR that is suggested by the partner's already-observed decision. The model selects the trial where the partner is most likely to do the opposite of what 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 Give; 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 Take.

The 'Confirmation' model does the opposite: it tries to select the trial where the partner is most likely to do the same thing as she did before. Just like the ideal search model, all three alternative models make choices in a stochastic manner.

Model fitting

All computational models are built on top of the ideal observer model, which must be equipped with a prior. Here I could not use the prior I had derived for study 2.2 in Chapter 2, because that study modeled people's inferences about decisions that cannot entail loss of money. In that study, the shape of the prior below WTRs of 0 did not matter much. In contrast, in the current experiment people sometimes have to think about decisions that entail loss of money (sometimes the dictator has the opportunity to lose \$15 to prevent the recipient from getting \$30), decisions for which negative WTRs are relevant.

Empirically, people's WTRs when playing the WTT as dictators show a sharp dis-

continuity at WTR = 0 (see figure 3.2). 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).

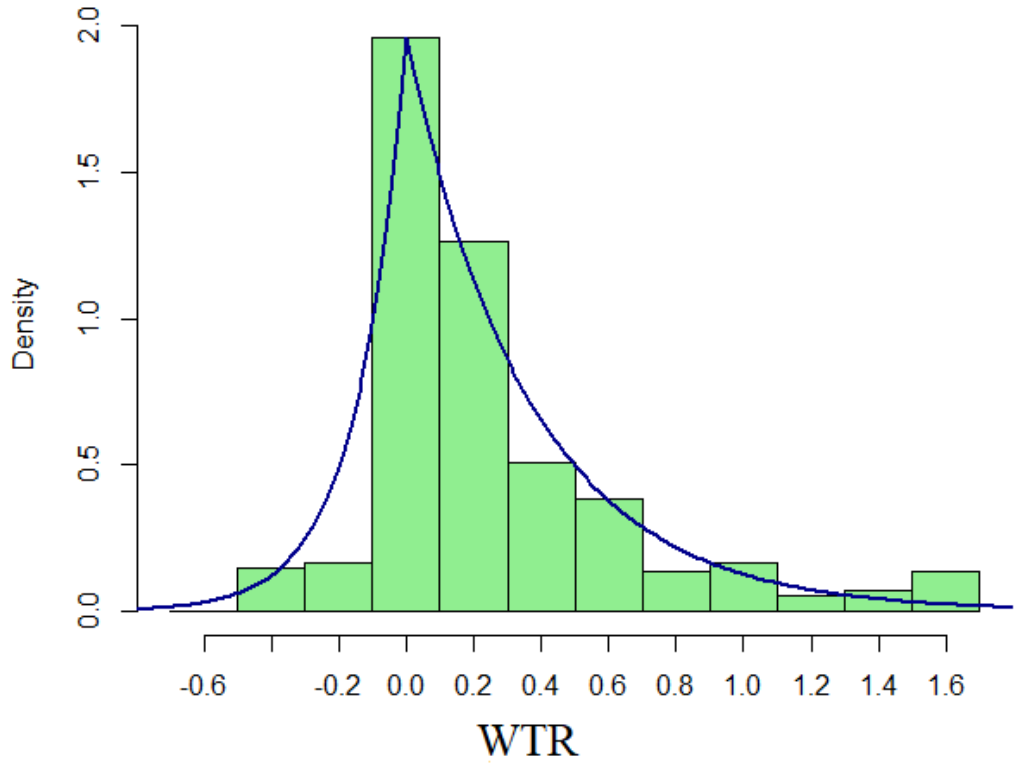


Figure 3.2: Histogram of WTRs in an MTurk sample (N=479, unpublished data; Sznycer et al.). In blue, the best-fitting skewed Laplacian distribution.

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 the ideal search model, I simultaneously fit these parameters by finding the values that minimized the root mean square error (RMSE) between model predictions and average human choices in the data selection task, using the optim function in R. Best-fitting values were skew=.23, dispersion=.37, β=2.22. Figure 3.3 plots the corresponding prior distribution (skewed Laplacian with location = 0, skew=.23, dispersion=.37).

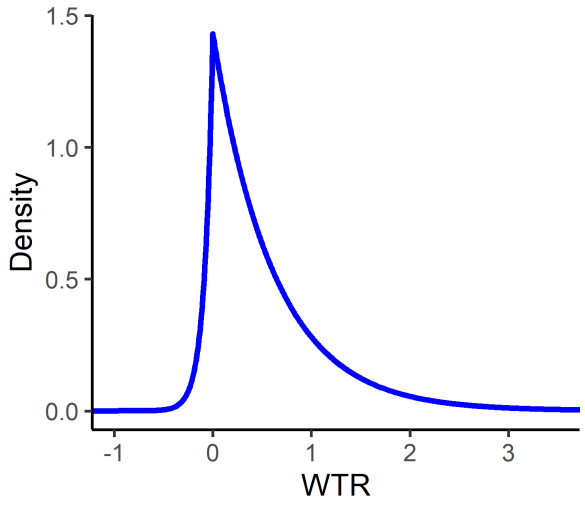


Figure 3.3: Prior belief of the ideal observer about the partner's WTR. Mean = .38, Standard Deviation = .44

To give a fair chance to alternative models, I tested two different parameter-fitting procedures for each one and kept the one that yielded the best fit. The first parameter-fitting procedure was simply the one described above. The second procedure consisted in using the prior inferred for the ideal search model, and optimizing the inverse temperature parameter alone. The second procedure worked best for the confirmation and falsification models, while the first one worked best for the 'optimal search without updating' model.

The data, and the R code for the computational model, data analysis and figures are available at the Open Science Framework[6].

[6]https://osf.io/bf6s4/?view_only=6b47266a55b847bab14a13f4d426292d

3.5 Results

Figure 3.4 displays the raw data, i.e. the average proportion of participants making a given choice for every pair of trials. Figure 3.5 plots the raw data for the choices made by ideal search model.

The first observation 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 participants} 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 $r(28) = .878$, $p < .001$. Figure 3.6 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. Figure 3.7 shows that the ideal

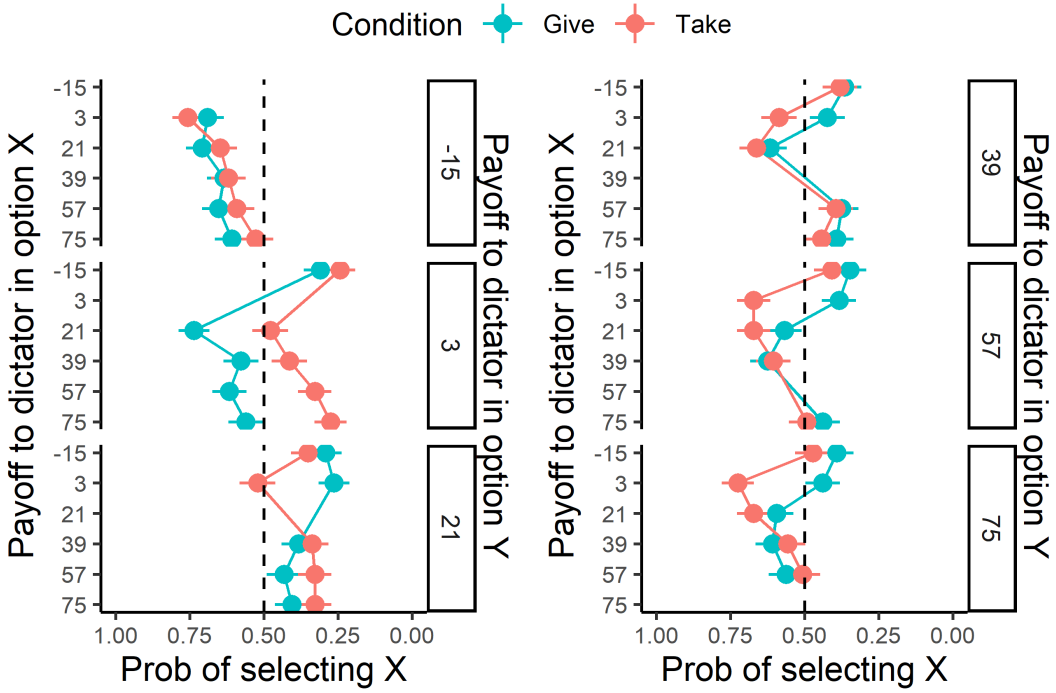


Figure 3.4: Average proportion of participants selecting a trial, for every possible pair of trials. Payoffs are in USD, and potential payoff to recipient was $30 in every trial. For instance, about 75% of people in the 'Give' condition selected the trial with {$21 for partner, $30 for participant} as 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 $\pi_{partner} = 3\}$.

search model will select trials with a higher value of $\pi_{dictator}$ 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 $\pi_{dictator}$ than participants in the "Take" condition; b = -4.5, p = .02; (linear mixed model with random intercepts, participants as random effect).

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

No. The choices of the ideal search model followed an inverted-U curve, and people's choices followed a similar pattern 3.7. To test for the statistical significance of this inverted-U curve pattern in the human data, I performed two-lines tests (Simonsohn,

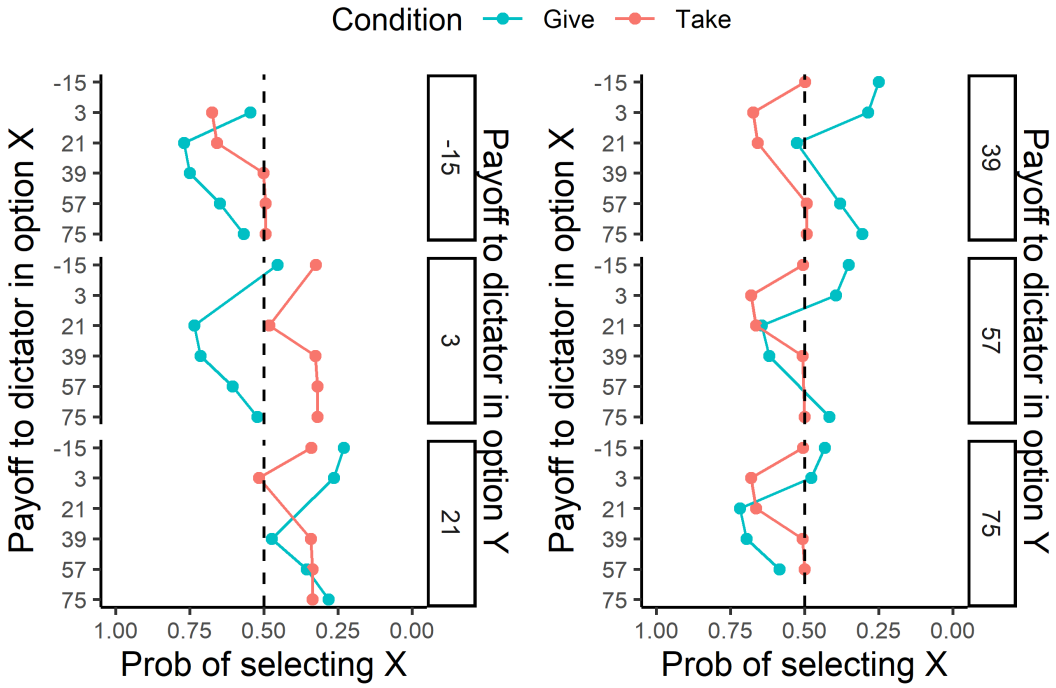


Figure 3.5: Probability that the stochastic ideal search model selects a trial, for every possible pair of trials. Payoffs are in USD, and potential payoff to recipient was $30 in every trial. For instance, in the 'Give' condition the ideal search model selected the trial with {$21 for partner, $30 for 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 $\pi_{partner} = 3\}$.

2018).

For participants in the Take condition, in the interval between $\pi_{dictator}$ = -15 and $\pi_{dictator} = 3$, the value of $\pi_{dictator}$ 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 $\pi_{dictator} = 3$ and $\pi_{dictator} = 75$, it was a negative predictor, b = -.02, p < .001.

For participants in the Give condition, in the interval between $\pi_{dictator}$ = -15 and $\pi_{dictator} = 21$, the value of $\pi_{dictator}$ in a trial was a positive predictor of the probability of selecting that trial; b = .05, p <.001 (multilevel logistic regression with random slopes

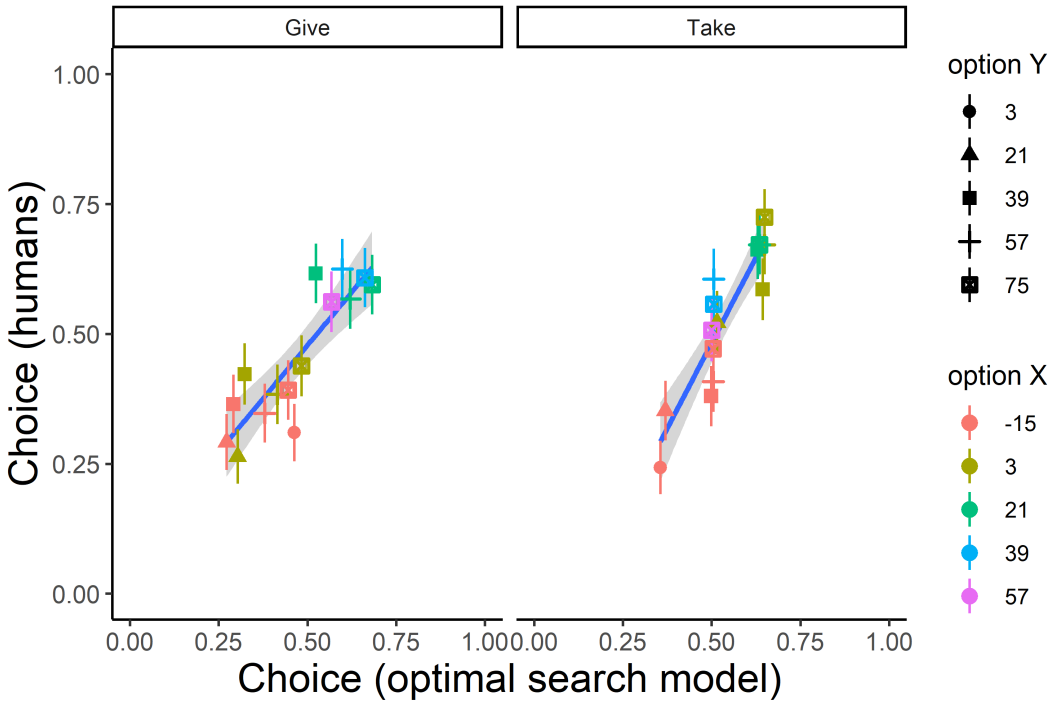


Figure 3.6: Relation between the choice probability of the optimal search model and the choice probability of human participants, presented separately for participants in the Give and the Take condition. Each point represents one pair of trials. Error bars represent the standard error of the mean. Higher values correspond to a higher probability of choosing option X.

and random intercepts, and participants as random effects). In the interval between $\pi_{dictator} = 21$ and $\pi_{dictator} = 75$, 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 figure 3.8. 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.

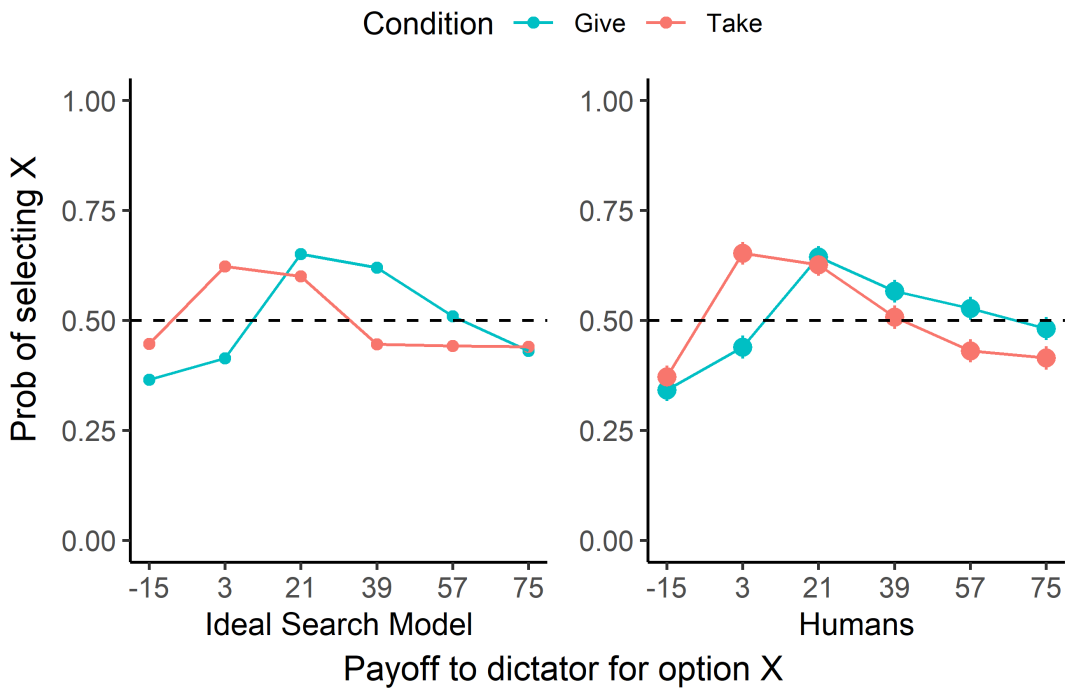


Figure 3.7: Probability of selecting a trial, as a function of potential payoff to dictator for that trial, for the ideal search model (left) and human participants (right). This graph collapses over all other potential values of $\pi_{dictator}$ for the other trial in the pair. Error bars represent the standard error of the mean.

The item-level correlation between people's choices and the choices made by the falsification model was r(28) = .235, p = .21; for the confirmation model, this correlation was negative, r(28) = -.235, p = .21.

I also compared the fits of the different models to the human data using a multilevel approach. For each model I computed a multilevel logistic regression with random slopes and intercepts, participants as random effects, participant choices as dependent variable, and model predictions as predictor. Table 3.1 shows the AIC score for each model (lower AIC scores indicate better fit to the data). Descriptively, the ideal search model had the best fit to the human data.

Paired permutation tests show that the ideal search model did not have a significantly better fit to the human data than the 'ideal search without updating' model, p = .197,

but had a significantly better fit than the confirmation and falsification models (both ps $< .001$).

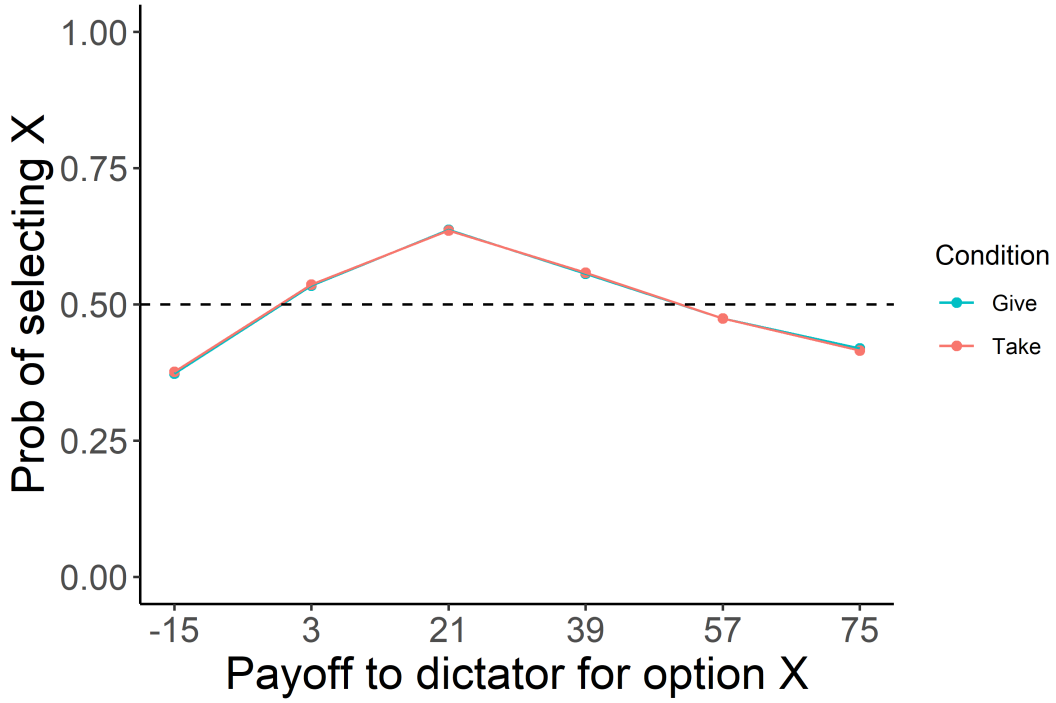


Figure 3.8: Probability that the ideal-search-without-updating model selects a trial, as a function of potential payoff to dictator for that trial. This graph collapses over all other potential values of $\pi_{dictator}$ for the other trial in the pair. The model makes the same predictions for both conditions because he is given the same information in both conditions.

Model	AIC	Pearson's r
Ideal Search	2731	.878***
Ideal Search No Updating	2755	.812***
Falsification	2748	.235 (n.s)
Confirmation	2748	-.235 (n.s)

Table 3.1: Akaike Information Criterion (AIC), and Pearson's correlation coefficient (r) for the fit of each search model to the human data. Lower AIC scores indicate better fit. AICs are derived from multilevel logistic regressions, while Pearsons rs are computed with simple correlation tests. ***: p $< .001$; n.s : p $> .05$

3.6 Study 3.2

Results of study 3.1 suggest that people spontaneously select the evidence that contains the most potential information about the WTR of their partner.

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 what payoff they would get, and selected the option in each pair for which a take versus give decision is most uncertain.

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). Indeed, I tested a search model that selects the trials with highest entropy, and found that it made the same predictions as the ideal search model based on KL divergence.

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.

Note that this account is not entirely deflationary: a causal model of how others make welfare trade-offs and the ability to make approximately Bayesian computations would still be necessary to compute which outcome is the most uncertain. The outcome-oriented 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).

In study 3.2, 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. By contrast, the trials where the partner makes the decision always have entropy lower than 50%. 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.

3.6.1 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).

3.6.2 Procedure

Study 3.2 was identical to Study 3.1, with the following exceptions. First, I omitted the 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 $\pi_{dictator}$ (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 3.1, 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.

3.6.3 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 $r(28) = .841$, $p < .001$. Figure 3.9 depicts the correlation between ideal search model and participant average choices, broken down by condition.

Figure 3.10 displays the raw data, i.e. the average proportion of participants making

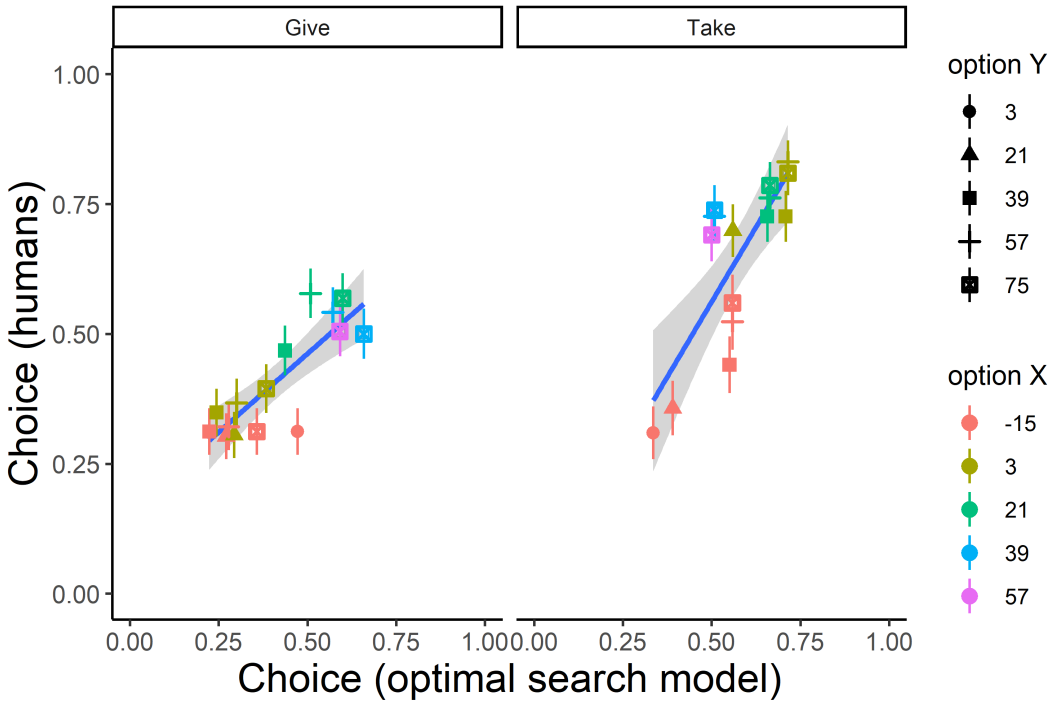


Figure 3.9: Relation between the choice probability of the optimal search model and the choice probability of human participants, displayed separately for participants in the Give and the Take condition. Each point represents one pair of trials. Error bars represent the standard error of the mean. Higher values correspond to a higher probability of choosing option X.

a given choice for every trial pair. Figure 3.11 plots the raw data 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. Figure 3.12 shows that the ideal search model will select trials with a higher value of $\pi_{dictator}$ 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 $\pi_{dictator}$ than participants in the "Take" condition; b = -10.6, p < .001; (linear mixed model with random intercepts,

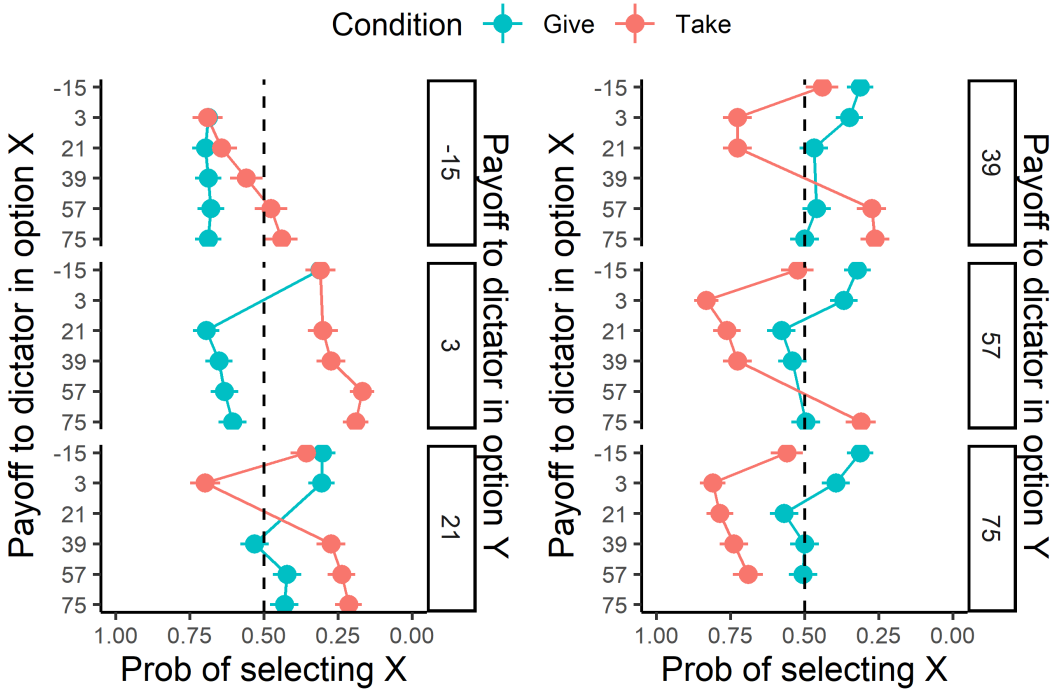


Figure 3.10: Proportion of human participants selecting a trial, for every possible pair of trials. Payoffs are in USD, and potential payoff to recipient was \$30 in every trial. For instance, in the 'Give' condition participants selected the trial with {\$21 for partner, \$30 for participants} over 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 $\pi_{partner} = 3\}$.

participants as random effect).

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

No. The choices of the ideal search model followed an inverted-U curve, and people's choices followed a similar pattern (Figure 3.12). 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 $\pi_{dictator}$ = -15 and $\pi_{dictator} = 3$, the value of $\pi_{dictator}$ in a trial was a positive predictor of the probability of selecting that trial; b = .10, p < .001 (multilevel logistic regression with random slopes

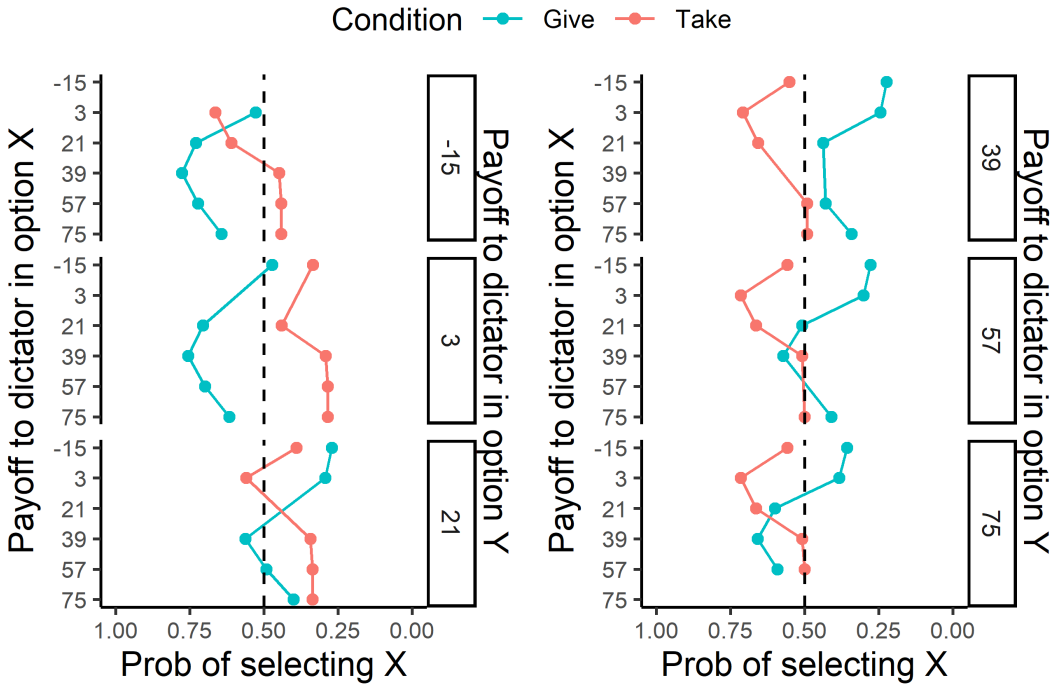


Figure 3.11: Probability that the stochastic ideal search model selects a trial, for every possible pair of trials. Payoffs are in USD, and potential payoff to recipient was $30 in every trial. For instance, in the 'Give' condition the ideal search model selected the trial with {$21 for partner, $30 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 $\{\pi_{partner} = 3$ vs $\pi_{partner} = 21\}$ is the same as the data point for $\{\pi_{partner} = 21$ vs $\pi_{partner} = 3\}$.

and random intercepts, and participants as random effects). In the interval between $\pi_{dictator} = 3$ and $\pi_{dictator} = 75$, it was a negative predictor, b = -.04, $p < .001$.

For participants in the Give condition, in the interval between $\pi_{dictator}$ = -15 to $\pi_{dictator}$ = 21, the value of $\pi_{dictator}$ 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 $\pi_{dictator}$ = 21 and $\pi_{dictator} = 75$, $\pi_{dictator}$ for a trial had no effect on the likelihood of selecting that trial , b = .00, p = .88.

In sum, participants were not consistently attracted to trials with extreme values of

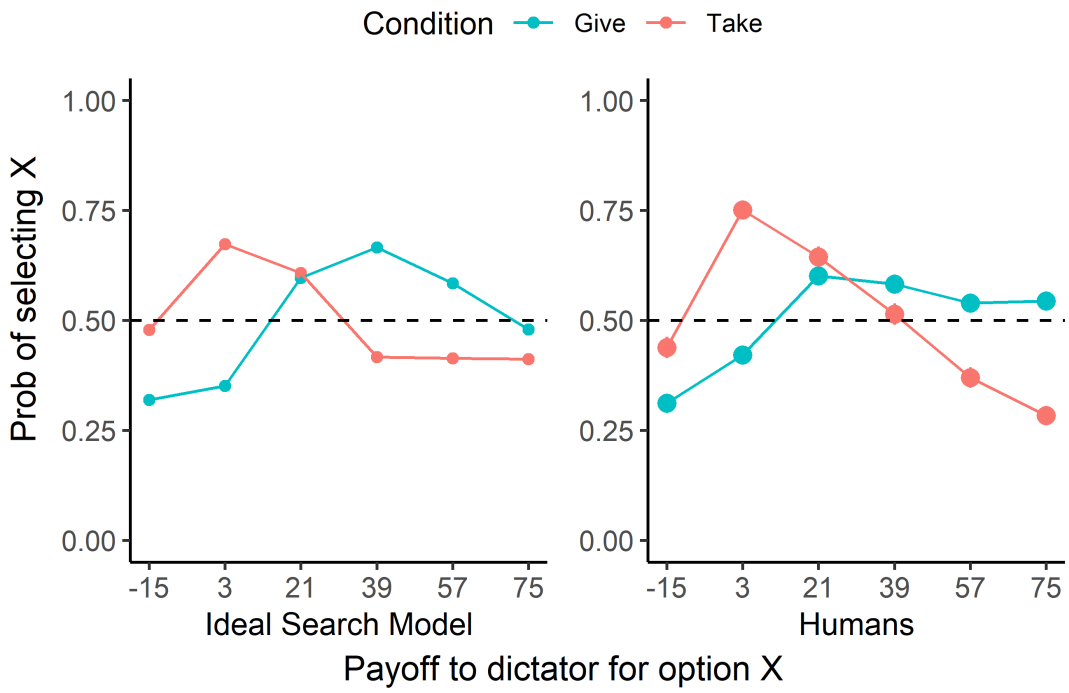


Figure 3.12: Probability of selecting a trial, as a function of potential payoff to dictator for that trial, for the ideal search model (left) and human participants (right). This graph collapses over all other potential values of $\pi_{dictator}$ for the other trial in the pair. Error bars represent the standard error of the mean.

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 $\pi_{dictator}$ the effect of $\pi_{dictator}$ 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) = -.468, p = .009.

I also compared the fits of the different models to the human data using a multilevel approach. For each model I computed a multilevel logistic regression with random slopes and intercepts, participants as random effects, participant choices as dependent variables, and model predictions as predictor. Table 3.2 shows the AIC score for each model (lower AIC scores indicate better fit to the data).

Finally, paired permutation tests show that the ideal search model had a significantly better fit to the human data than all other models, all ps $< .001$.

Model	AIC	Pearson's r
Ideal Search	3539	.841***
Ideal Search No Updating	3665	.635***
Falsification	3610	.520**
Confirmation	3612	-.468**

Table 3.2: Akaike Information Criterion (AIC), and Pearson's correlation coefficient (r) for the fit of each search model to the human data. Lower AIC scores indicate better fit. AICs are derived from multilevel logistic regressions, while Pearsons rs are computed with simple correlation tests. ***: p < .001; **: p < .01

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.

Figure 3.13 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.

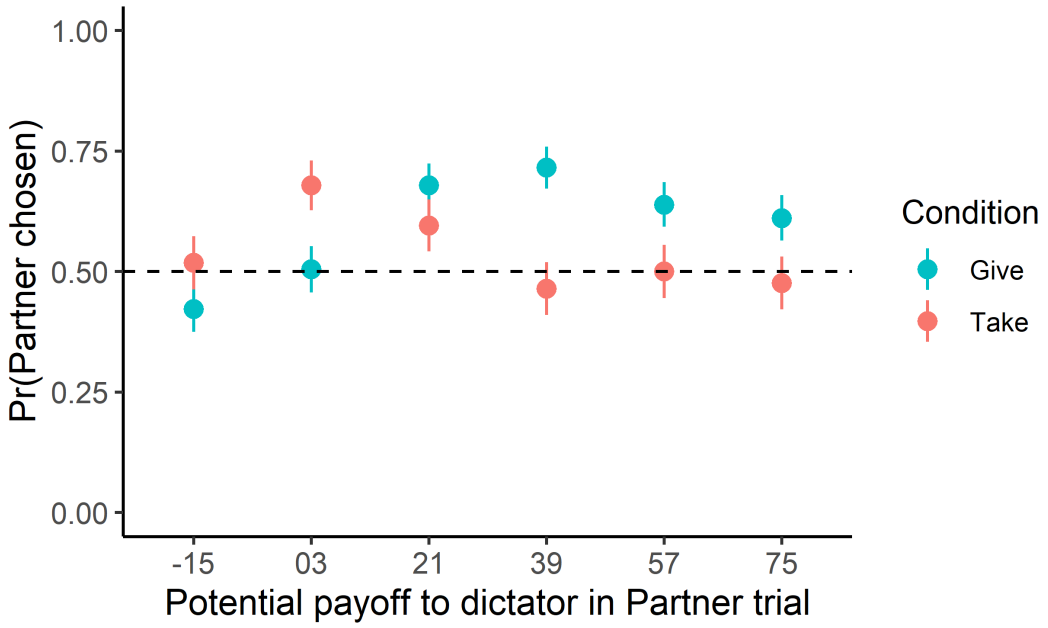


Figure 3.13: Proportion of participants who selected the partner-determined trial instead of the computer-determined trial, as a function of the potential payoff to the partner. Error bars represent the standard error of the mean.

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

3.6.4 Study 3.2 Discussion

Study 3.2 replicates the main results of Study 3.1: 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 the output of a psychology designed to extract information about a causally deep property of the social world: someone's valuation of your welfare.

3.7 Discussion

In a simple data selection task, the trial with the greatest potential to reveal their partner's WTR elicited more curiosity from participants. Participants preferred those options even though they were not given any explicit criterion for how to make their choices: participants were simply asked which data they would most like to observe. Furthermore, I did not give any information that might have suggested to participants that they would need to infer the weight that their partner put on their welfare[7].

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 they had previously seen their partner act selfishly or generously. 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.

Many studies on human data selection give participants explicit information about the goal of the task and about the information structure of the task. 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

[7]One might argue that the prediction task that participants also completed might have primed them to look for WTR-relevant information. I tested this 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.

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, the current task requires the possession of domain-specific knowledge about how people typically make welfare trade-offs. Therefore its results show that people are able to spontaneously mobilize their domain-specific causal knowledge in order to guide information-gathering.

The current task was very simple: different trials of the WTT differ only in the potential payoff for the dictator. This feature of the design was necessary in order to make the construction of the ideal search model tractable. Future research should investigate the extent to which data selection about social valuation is sensitive to information value in more complex settings, where other parameters vary (for instance, benefit to the participant, intentionality).

The question remains whether people would prioritize information about WTR to information about other traits (attractiveness, competence, overall generosity, etc). As mentioned in chapter 1, Wojciszke et al. (1998) found preliminary support for this: when asked which traits they most would 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*. It would be interesting to replicate this finding in a more formal setting, for instance by giving participants potential queries that differ in their information content with respect to competence and WTR, and test whether (and to what extent) people privilege WTR-relevant information.

In sum, 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 provides more evidence that the mind houses cognitive machinery that models the welfare-tradeoff behavior of others. More generally, it also supports the idea that human information search is efficient when the task activates the right inference systems.

Chapter 4

Discussion and conclusions

A growing body of research suggests that the regulation of welfare tradeoffs plays a fundamental role in social cognition. The mind has mechanisms that adjust the tradeoffs that one makes between one's welfare and that of specific others; represent the disposition of others to make such tradeoffs; and regulate the welfare tradeoffs made by others. This theoretical framework, emerging from evolutionary biology, has been supported by empirical evidence that people make inferences about the Welfare Tradeoff Parameters (WTPs) of others, and that these inferences regulate emotion and behavior.

The current dissertation contributes to this body of research by investigating how people make inferences about the WTPs of others. In chapter 2, I show that a normative model of inference under uncertainty is able to account for people's inferences about the WTPs of others in a simple prediction task. These inferences appear to be an input to anger, and possibly gratitude. The chapter also presents tentative evidence that the surprise (in an information-theoretic sense) elicited by an observation may intensify the magnitude of the emotion elicited by this observation. In chapter 3, I provide further evidence that people efficiently extract WTP-relevant information from their environment, by showing that they tend to request to observe the events that contain the most

potential information about the WTPs of others.

By showing that humans have the necessary machinery to efficiently construct and update their representations of the WTPs of others, these findings support the hypothesis that these representations play an important role in human social cognition. At the same time, they also provide evidence for a specific view of the origins of human (ir)rationality. According to that view, human reasoning is supported by a myriad of specialized inference mechanisms; humans are expected to solve a problem rationally provided that the problem is in a format that activates the relevant inference mechanism. Here, the adaptive importance of social valuation inference led to the hypothesis that the human mind houses mechanisms that recognize a social valuation inference problem as such, and solve this problem near-optimally.

In this chapter, I discuss some of the implications of this work and point to possible directions for future research. First, I discuss what is and is not meant by the claim that human behavior conforms to an 'optimal' model. Then I discuss the ontogeny of social valuation inference: how does the ability develop in humans, and is it the product of mechanisms that also solve other types of inference problems? Later sections discuss prospects for a more detailed investigation of social valuation inference, including the prospect for a more holistic model of welfare trade-off cognition that models decisions and inferences simultaneously.

4.1 What do I mean by 'optimality'?

I claimed that human performance in the current studies can be reasonably-well modeled by assuming that participants are solving the task in the mathematically optimal way. To clarify what this means, it is important to make two distinctions. The first is the distinction between optimality at the level of design vs individual performance. The

second is the distinction between levels of analysis in cognitive science.

4.1.1 Design vs individual performance

The current experiments were designed to test for optimality at the level of design. They test the hypothesis that the reliably-developing mechanisms underlying social valuation inference are designed such that, under the right conditions, they are able to solve an inference problem in (approximate) conformity with normative standards of probabilistic inference, with the help of the appropriate domain knowledge.

This is not equivalent to testing whether individual human beings are systematically optimal at this task. Indeed, a quick glance at the plots for the individual-level data in Chapter 2 (see section 2.6) reveals some variability in human performance: while some individuals come very close to ideal-observer predictions, others are far off the mark. Evidence for good design is assessed by looking at whether, on average, people tend to make inferences in the normatively correct way. Even assuming (near-) optimality at the level of cognitive design, at the individual level one expects a variety of factors (noise in information-processing, slips of the mouse, inattention to the task, etc) to sometimes cause sub-optimal performance.

4.1.2 Levels of analysis

The current studies are mostly at what David Marr (1982) called the computational level of analysis. The ideal observer models used here are designed as a theory of the computation that the mind is solving in the tasks. I do not claim that they specify the exact algorithms that people use – such a claim would belong to Marr's algorithmic level of analysis.

At the algorithmic level, there are good reasons to think that humans do not run

exact Bayesian computations when solving social valuation inference tasks. People have to make such inferences in settings vastly more complex than the simple experiments used here. In these settings, Bayesian inference is likely to be computationally intractable. Therefore an exhaustive theory of these inferences would need the tools of the literature on 'bounded' rationality (Simon, 1955). Recent work in cognitive science suggests that one may build such algorithmic-level theories of cognition by looking for efficient ways to approximate the computations specified by computational-level theories (Griffiths, Vul & Sanborn, 2012; Lieder & Griffiths, 2020).

4.2 The ontogeny of social valuation inference

Under the current proposal, humans infer how much others value them by making probabilistic computations over a causal model of the way agents typically make welfare trade-offs. The abstract concept of Welfare Tradeoff Parameters plays an important role in that causal model. By what mechanisms do people acquire this concept? Although the current data, collected on American adults, do not speak directly to that question, here I explore some possibilities.

Humans have a set of reliably-developing mechanisms that allow them to reason about the minds of others (Leslie, German & Friedman, 2004; Onishi & Baillargeon, 2005; Barrett et al., 2013). It is likely that these mechanisms play an important role in our ability to reason about the WTPs of others. Recent work suggests that many of the inferences people make about the mind and behavior of others relies on Bayesian inference over causal models of others' minds (Baker et al., 2009, 2017; Lucas et al., 2014; Jara-Ettinger et al., 2016). Thus, a lot of the machinery that allows us to make welfare-tradeoff inferences might belong to the reliably-developing mechanisms underlying commonsense psychology – this suggests that they may appear early in development,

and cross-culturally.

Could it be, then, that we learn the concept of welfare trade-off parameters just as we learn about other preferences? For instance, even if people do not have innate knowledge of the concepts of 'apples' and 'oranges', they can learn that it is possible to prefer apples to oranges. By the same token, maybe the human mind has no genetically encoded knowledge about welfare trade-offs, but we learn by observation that people sometimes behave as if they place some value on the welfare of others. Assuming that concepts such as 'costs' and 'benefits' are conceptual primitives in commonsense psychology (Jara-Ettinger et al., 2016), then the child may learn that people sometimes incorporate costs and benefits to others in their decision-making. Under such an account, social valuation inference is a straightforward byproduct of commonsense psychology.

Though parsimonious, this account would have difficulty accounting for the interplay between welfare tradeoff inference, emotion and motivation. For instance, 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; 2018), and feel shame in response to traits and events that make others less likely to value them (Sznycer et al., 2016; 2018). Social valuation inferences also strongly motivate us to recalibrate our own WTPs (Smith et al., 2017; Lim, 2012) toward others. If WTP-related concepts are not genetically encoded, why do cues of social valuation trigger similar emotional reactions in so many different cultures?

From a theoretical point of view, one expects motivational systems to co-evolve with their proprietary concepts (Tooby, Cosmides & Barrett, 2005; Delton & Sell, 2014; Barrett, Cosmides & Tooby, 2010; Cosmides, Guzman & Tooby, 2018). A fear of predators is of no use to an organism that does not have a concept of predator (Barrett, 2015). A motivation to help siblings and avoid mating with them cannot guide behavior unless one has a concept of sibling (Lieberman, Tooby & Cosmides, 2007). If humans simply

learned the concept of sibling via domain-general learning mechanisms, in the absence of any genetically encoded information, the motivational system would not 'know' that this concept (as opposed to thousand others) should regulate one's altruism and mating behavior in a highly specific way.

Successful social interaction in humans relies on evolved motivational systems that lead us to seek the company of those who value us, deliver benefits (or inflict costs) contingent on cues of a person's valuation of us, send signals that we expect to be valued more, etc. For these systems to have evolved, there needed to be a tight mesh between the behavior-regulating machinery (i.e. motivation) and concepts of social valuation. Without an evolved set of concepts related to social valuation, evolution could not have designed decision rules for social interaction that take social valuation information as input.

The necessary coevolution of concepts of motivation does not imply that the concepts over which motivational systems operate spring fully-formed at birth, independently of any experience. What it does imply is that the organism has at least some embryological version of the concept, ready to imbue newly acquired information with motivational relevance.

For instance, one can imagine that a relatively general system within commonsense psychology, whose task is to infer the preferences of others, allows us to infer the WTPs of someone else towards us. Under this scenario, we infer "Alice values my welfare half as much as she values hers" by the same mechanisms that allows us to infer "Alice likes apples half as much as she likes oranges". But even then, there would need to be some downstream system that scrutinizes the preferences inferred by this inference mechanism and tags them as relevant for systems that compute anger, gratitude, etc. This scrutiny system would need to be equipped with a built-in template enabling it to categorize the output of a preference inference as being a preference over welfare tradeoffs.

More realistically, it is likely that the inference systems themselves exhibit at least some degree of specialization for social valuation inference. For instance, upon observing that Alice decides to give us a slice of cake, we need to infer something more general than "Alice values the fact that we have cake". In other words, we need to represent Alice as having preferences about the relatively abstract notions of costs and benefits to us. One can speculate that a tendency to represent agents as meta-representing the valuation systems of others is a reliably-developing feature of the human mind, which considerably speeds up the development of social valuation inference. Tentative evidence for this hypothesis comes from studies that suggest that infants categorize agents according to whether they hinder or help another agent's goal (Hamlin, Bloom & Wynn, 2007; but see Schlingloff, Csibra & Tatone, 2020). Evidence that infants have an abstract concept of giving (Tatone, Geriaci & Csibra, 2015; Tatone, Hernik & Csibra, 2019) suggests specialized machinery to reason about abstract features of resource transfer, which could spur the development of welfare tradeoff reasoning.

In sum, a plausible account is that mechanisms for commonsense psychology, in combination with more specialized representational systems, underlie the development of social valuation inference in humans.

4.3 Directions for future research

In settings more complex than the simple tasks explored here, are people's inferences still reasonable? Research could profitably investigate human performance in inference problems with more parameters (e.g. uncertainty about intentionality). At an algorithmic level, more complex tasks could also reveal the mechanistic basis of inference about social valuation (for instance, what – if any – heuristics do people use?). It is also notable that in the current work I have only studied inference about a single parameter

(the WTR). Future research could test inferences in contexts where human welfare trade-offs are best modeled with utility functions that have more than one WTP (for instance, weight assigned to other's welfare and elasticity of substitution, Andreoni & Miller, 2002; Fisman, Markovits & Kariv, 2007).

Speculations in the previous section (4.2) could be evaluated with developmental research that looks at whether infants and young children spontaneously represent social events in terms of welfare trade-offs. For instance, do young children make general or specific inferences about generosity (if Alice shared her cake with Bob, do they infer that she would also give him some stickers in a later dictator game)? When watching a helping act, do they use costs to the actor as a cue to the actor's valuation of the recipient?

Exploratory findings from chapter 2 provide tentative evidence for the role of surprise (i.e. deviation from expectation) in anger and gratitude. These results are consistent with previous findings that people's expectations about the amount of money they will be offered in an ultimatum game predict the threshold at which they will reject offers, and their self-reported happiness at the offer (Sanfey, 2009; Chang & Sanfey, 2011; Xiang, Lohrenz & Montague, 2013). However, future work is needed to ensure that the current findings are not a measurement artifact. For instance, one could experimentally manipulate the prior expectations of participants, by showing them either generous or selfish decisions before the main task (as in Xiang, Lohrenz & Montague, 2013).

If these results are conceptually replicated, they would open up opportunities for deeper theorizing about how to construe 'expectations'. Expectations may be purely statistical, but they may also take a more prescriptive meaning: for instance someone may feel like they deserve to be treated well even though they believe this is unlikely to happen. Colloquially, when we say "I expect to be treated well", we are often bargaining for good treatment instead of making a descriptive statement. Let us call 'entitlement' this non-statistical meaning of expectation (e.g. Sell et al., 2009). Correlational studies

of individual differences suggest a role of entitlement in anger (Sell et al., 2009), but there is no accepted model of how entitlement is represented in the mind. Could it be conceptualized as a curve, with lower revealed WTPs represented as increasingly unacceptable? If so, what determines the steepness of that curve? How does this curve change as the individual receives new information about, e.g., relative formidability or outside options of the interactants?

4.3.1 Potential for a recursive model of welfare-tradeoffs

Although the current work focuses on inferences about welfare-tradeoffs, it has potential implications for how people may make welfare-tradeoffs in the first place.

In the modeling framework I used here, Alice has a utility function which determines how she trades off Bob's welfare against hers, and Bob tries to infer the relevant parameters of that utility function. Thus, the utility function is primary, and the inference model is built on top of it. However, there are theoretical and empirical reasons to think the relationship goes both way. Alice's WTPs toward Bob may depend in great part on her inferences about Bob's WTPs toward her, as well as the inferences she wants Bob to draw about her WTPs (Tooby & Cosmides, 1996; Lim, 2012). This recursive character of social valuation is not currently captured by the simple modeling framework that I use here, and is not captured by economic theories of "social preferences" in general (e.g. Fehr & Schmitt, 1999; Charness & Rabin, 2002).

To give a concrete example of the problem, consider a study by Bardsley (2008; see also List, 2007). The author conducted a standard dictator game, as well as a 'taking game', a slightly modified dictator game where dictators have the same options as in the standard game, except that they can also take money from the recipient. Standard models of social preferences predict that the distribution of positive allocations should be

similar across experiments (i.e. the proportion of dictators giving $1 should be identical, the proportion of dictators giving $2 should be identical, etc, although the proportion of people giving $0 may differ). This is not what Bardsley observed: significantly fewer people gave money in the taking game compared to the standard dictator game. This result can be interpreted in terms of participants accounting for the inferences that can be drawn from their behavior. In a standard dictator game, giving $0 is the most selfish option you can take, and so giving $0 establishes no lower bound in the eyes of others on your potential selfishness. This creates an incentive for people to give some amount, so that such a lower bound can be established in the minds of onlookers. By contrast, in the taking game giving $0 is not the most selfish action you can take (nor is, for that matter, taking half of what you could potentially take from the recipient's endowment), so it is a more attractive option.

Such results (see also List, 2007; Dana, Cain & Dawes, 2006; Dana, Weber & Kuang, 2007; Burum, Hoffman & Nowak, 2020) suggest that a formal approach to welfare tradeoffs needs to incorporate a model of the inferences that the agent thinks will be made by the audience[1].

How would one build such a model? A potentially helpful analogy is that of human communication. The pragmatics research program in linguistics (Grice, 1975; Sperber & Wilson, 1986) emphasizes that communication involves a reconstructive process, where the listener infers the meaning of the speaker on the basis of the sparse evidence contained in the speaker's utterance. But because the speaker's goal is to produce such an inference in the mind of the listener, she needs to speak in a way that accounts for the inferences that she thinks the listener is going to make. In turn, the listener expects that the speaker

[1]Of course, this does not imply that there needs to actually be an audience, or that the agent is making (consciously or unconsciously) meta-inferential computations. One simply expects that the machinery for welfare tradeoffs is well-designed to make the kinds of decisions that enhance (or do not excessively damage) the agent's reputation.

speaks in a way that accounts for the inferences he will make.

Here is a concrete example. Alice and Bob both win $1 if Bob selects the correct shape among the three shown in Figure 4.1. Alice knows that the correct shape is the blue square, but she can only communicate this to Bob by saying either "blue" or "square" – what should she say?

Figure 4.1: Stimulus in a simple language game.

Neither word unambiguously picks a single shape, so at first sight it may seem that both are equally informative. However, Alice and Bob can ensure that they win the prize if they recursively model each other's minds. If the correct shape was the green square, then Alice could say "green" and unambiguously refer to the green square. Knowing this, if Bob hears anything other than 'green', he knows that the correct shape is not the green square (because if it was, Alice would have said 'green'). Therefore Alice knows that if she says "blue" or "square", Bob will infer that the correct shape is not the green square. Therefore, by saying "square", she can unambiguously refer to the blue square. Empirically, people solve these sorts of simple language games in conformity with such a recursive inference model (Frank & Goodman, 2012).

Welfare tradeoffs are slightly different, in the sense that the inferences that the audience draws are not the only thing that the decision-maker is trying to optimize. Still, because these inferences are plausibly a constraint on the decision-makers' welfare-tradeoff decisions, the recursive Bayesian approach may allow one to build elegant models of welfare-tradeoffs (by simultaneously modeling decision and inference) which could account for human behavior in the Bardsley (2008) study and similar experiments.

Such work may have important practical implications. Under such a model of welfare tradeoffs, the inferential consequences of a given action may matter more to the decision-maker than its material consequences for the beneficiaries. Theoretical models of 'warm-glow' altruism in economics (Andreoni, 1990), and data from psychology and experimental economics (Burum, Hoffman & Nowak, 2020; Dana, Cain & Dawes, 2006; Dana, Weber & Kuang, 2007) highlight that people often optimize their altruistic decisions toward making themselves feel good rather than having an effective material impact on the recipients of the altruistic act. A good theoretical model of why this is the case would improve the prospects of efforts to make human altruism more focused on concrete impact (Singer, 2015).

It is likely that "warm glow", the good feeling one gets from performing an altruistic action is an internal signal of the good inferential consequences of one's act (i.e. how good one looks in the eyes of the audience as a result), in the same way that pain is an internal signal of bodily damage, and pride is an internal signal of the increased valuation one gets from a given trait or achievement (Sznycer et al. 2017). Supporting this view, Yudkin, Prosser & Crockett (2019) found that the warm glow participants reported they would feel if performing a given act was strongly correlated with the praiseworthiness of that act, as judged by a separate sample of participants, independently of the magnitude of the actual good caused by the act[2].

It might be possible to conduct a more formal test of this hypothesis. For instance, ask participants to imagine that they play a lottery game where, if they win, someone else earns $X, but if they lose, they themselves lose $Y. In all conditions, tell participants that the game is secretly rigged so that they always win. Holding X constant, then Y should positively predict participants' reports of warm glow, and the praiseworthiness

[2]Although Yudkin et al. have a different interpretation of their findings, not grounded in adaptationist principles.

ratings from a separate sample, because a high potential cost warrants more favorable inferences about the participants' WTPs.

4.4 Conclusion

The theory of evolution by natural selection, and the computational theory of the mind, jointly explain why the mind exists and how it works (Pinker, 1997). In this dissertation, I use the conceptual tools of both theories in complementary ways. Theories of social emotion and motivation, derived from evolutionary models of cooperation and conflict, specify a certain class of representations that the mind is predicted to use when parsing the social world. Theories of inference, derived from statistics and computer science, specify how a well-designed organism is predicted to efficiently construct and update such representations. On their own, statistical theories of inference are powerless for specifying the inferences that the organism should draw: they cannot specify a priori the appropriate background knowledge that enables fast learning from sparse data or the representations that these inferences should produce. On their own, the evolutionary models cannot make quantitative predictions about the inferences drawn by the organism or the kinds of evidence that it will look for. Jointly, they allow the derivation of rich, quantitative, and empirically testable predictions about inferences, emotions, curiosity, motivation, and the complex functional relationships between them.

Appendix A

Noise parameter for the ideal observer

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 $\sigma_\phi = .16$.

Appendix B

First parametrization of the prior used in chapter 2

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

Appendix C

Second parametrization of the prior used in chapter 2

One potential downside of the first parametrization is that I estimated the distribution of WTRs in MTurk participants, which may not exactly match the population that comes to the mind of participants when they think of their acquaintances. Another potential concern is that the Welfare-Tradeoff Task was probably new to most participants, and even if we assume that they have a good generative model of how people typically behave when they make welfare trade-offs, there is no strong reason to expect them to have perfectly accurate priors for that specific task. Therefore, for the second parametrization of the prior, I directly attempted to infer the prior belief that participants had about the distribution of WTRs among their acquaintances. I did so by asking participants in study 2 to complete a preliminary task, at the beginning of the experiment, where they had to make predictions about the behavior of interaction partners in the WTT for which they had not had an opportunity to see any other WTT decision (see chapter 2, section2.4 for details). For this part of the analysis, I only analyzed data from participants who exhibited a negative correlation between the cost of giving in a trial ($\pi_{dictator}$) and

the participant's prediction for that trial during this preliminary prediction task. A participant's failure to predict that the likelihood of giving gets lower as the cost of giving increases suggests that the participant was not paying attention to the task. 8 participants did not meet the criterion, and thus did not provide interpretable data.

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

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

Appendix D

Information-theoretic measure of surprise, chapter 2

The information-theoretic measure of surprise I use in chapter 2 is the Kullback-Leibler divergence. It is the same measure I use to quantify the information value of an observation when deriving the ideal search model in chapter 3. The KL divergence measures how much your probability distribution over the space of possible hypotheses shifts in response to the observation. Formally it is defined as:

$$KL = \sum_i P(h_i|d) \log\left(\frac{P(h_i|d)}{P(h_i)}\right)$$

Where $P(h_i)$ is the prior probability you assign to hypothesis h_i, and $P(h_i|d)$ is your posterior belief in h_i given the observation d.

In the context of WTR, because the hypothesis space is continuous, when formally writing the formula for KL divergence we use an integral sign instead of a summation

sign:

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

In order to compute the surprise that a given participant would experience when witnessing the pair of decisions made by a partner, I first computed, for that participant, an estimate of the prior of that participant about the WTR of an acquaintance (see Appendix B.). I then created an ideal observer with the same prior as the participant. I then computed the posterior belief of this ideal observer about the WTR of the partner, upon observing the decisions made by that partner. Finally, I computed surprise as the KL divergence between that posterior and the prior.

Bibliography

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

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The economic journal, 100*(401), 464-477.

Andreoni, J., & Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica, 70*(2), 737-753.

Ariely, D. (2008). *Predictably irrational*. New York, NY: Harper.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211*(4489), 1390-1396.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition, 113*(3), 329-349.

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.

Barclay, P. (2013). Strategies for cooperation in biological markets, especially for humans. *Evolution and Human Behavior, 34*(3), 164-175.

Bardsley, N. (2008). Dictator game giving: altruism or artefact?. *Experimental Economics, 11*(2), 122-133.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*(3), 211-233.

Baron, J. (1985). *Rationality and intelligence*. Cambridge University Press.

Barrett, H. C. (2005). Enzymatic computation and cognitive modularity. *Mind & language, 20*(3), 259-287.

Barrett, H. C., Cosmides, L., & Tooby, J. (2010). Coevolution of cooperation, causal cognition and mindreading. *Communicative & integrative biology, 3*(6), 522-524.

Barrett, H. C., Broesch, T., Scott, R. M., He, Z., Baillargeon, R., Wu, D., ... & Laurence, S. (2013). Early false-belief understanding in traditional non-Western societies. *Proceedings of the Royal Society B: Biological Sciences, 280*(1755), 20122654.

Barrett, H. C. (2014). *The shape of thought: How mental adaptations evolve*. Oxford University Press.

Barrett, H. C. (2015). Adaptations to predators and prey. *The handbook of evolutionary psychology*, 1-18.

Baumard, N., André, J. B., & Sperber, D. (2013). A mutualistic approach to moral-

ity: The evolution of fairness by partner choice. *Behavioral and Brain Sciences, 36*(1), 59.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical transactions of the Royal Society of London*, (53), 370-418.

Biernaskie, J. M., Walker, S. C., & Gegear, R. J. (2009). Bumblebees learn to forage like Bayesians. *The American Naturalist, 174*(3), 413-423.

Boyer, P., & Barrett, H. C. (2015). Intuitive ontologies and domain specificity. *The handbook of evolutionary psychology*, 1-19.

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

Burum, B., Nowak, M. A., & Hoffman, M. (2020). An evolutionary explanation for ineffective altruism. *Nature Human Behaviour*, 1-13.

Casscells, W., Schoenberger, A., & Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine, 299*(18), 999-1001.

Chang, L. J., & Sanfey, A. G. (2013). Great expectations: neural computations underlying the use of social norms in decision-making. *Social cognitive and affective neuroscience, 8*(3), 277-284.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The quarterly journal of economics, 117*(3), 817-869.

Clutton-Brock, T. H., & Parker, G. A. (1995). Punishment in animal societies. *Nature, 373*(6511), 209-216.

Cosmides, L. M., & Tooby, J. (1981). Cytoplasmic inheritance and intragenomic conflict. *Journal of theoretical biology, 89*(1), 83-129.

Cosmides, L. (1985). *Deduction or Darwinian algorithms. An explanation of the "elusive" content effect on the Wason selection task.* Unpublished doctoral dissertation, Harvard University.

Cosmides, L. (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. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58*(1), 1-73.

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.

Cosmides, L., Barrett, H. C., & Tooby, J. (2010). Adaptive specializations, social exchange, and the evolution of human intelligence. *Proceedings of the National Academy*

of Sciences, 107(Supplement 2), 9007-9014.

Cosmides, L., Guzmán, R. A., & Tooby, J. (2018). The evolution of moral cognition. In *The Routledge handbook of moral Epistemology*. Routledge.

Cox, R. (1961). *The algebra of probable inference*. Baltimore, MD: Johns Hopkins University.

Dana, J., Cain, D. M., & Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and human decision Processes, 100*(2), 193-201.

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory, 33*(1), 67-80.

Dawkins, R. (1976). *The selfish gene*. Oxford university press.

Dawkins, R. (1982). *The extended phenotype*. Oxford University Press.

Debove, S., Baumard, N., & André, J. B. (2017). On the evolutionary origins of equity. *PLoS one, 12*(3), e0173636.

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

Delton, A. W., & Robertson, T. E. (2012). The social cognition of social foraging:

Partner selection by underlying valuation. *Evolution and human behavior, 33*(6), 715-725.

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

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

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.

Eisenbruch, A., & Krasnow, M. (2019). *Why warmth matters more than competence: new evolutionary models.* PsyArxiv

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*(6870), 429-433.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics, 114*(3), 817-868.

Fessler, D. M., & Haley, K. J. (2003). The strategy of affect: Emotions in human cooperation. In *The Genetic and Cultural Evolution of Cooperation*, P. Hammerstein, ed, 7-36.

de Finetti, B. (1931). Sul significato soggettivo della probabilita. *Fundamenta mathematicae, 17*(1), 298-329.

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.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science, 336*(6084), 998-998.

Friedman, M. (1953). The methodology of positive economics. In *Essays in positive economics*. University of Chicago Press.

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural computation, 24*(1), 1-24.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological review, 102*(4), 684.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41-58). Brill.

Griffiths, T. L., & Tenenbaum, B. (2001). Randomness & Coincidences: Reconciling Intuition and Probability Theory. In *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society* (p. 370). Lawrence Erlbaum Associates.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological science, 17*(9), 767-773.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science, 21*(4), 263-268.

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.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge university press.

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.

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*, 103948.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature, 450*(7169), 557-559.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. *Journal of*

theoretical biology, 7(1), 17-52.

Hammerstein, P., & Parker, G. A. (1982). The asymmetric war of attrition. *Journal of Theoretical Biology, 96*(4), 647-682.

Hayden, B., & Niv, Y. (2020). *The case against economic values in the brain.* PsyArxiv

Heider, F. (1958). *The psychology of interpersonal relations.* Psychology Press.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge university press.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review, 90*(4), 293.

Kahneman, D. (2011). *Thinking, fast and slow.* Macmillan.

Kelley, H. H. (1973). The processes of causal attribution. *American psychologist, 28*(2), 107.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental science, 10*(3), 307-321.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55*, 271-304.

Kiers, E. T., Duhamel, M., Beesetty, Y., Mensah, J. A., Franken, O., Verbruggen, E., ... & Buecking, H. (2011). Reciprocal rewards stabilize cooperation in the mycorrhizal symbiosis. *Science, 333*(6044), 880-882.

Knill, D. C., & Richards, W. (Eds.). (1996). *Perception as Bayesian inference*. Cambridge University Press.

Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General, 136*(3), 430.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics, 22*(1), 79-86.

Laplace, P. S. (1812). *Theorie analytique des probabilites*. Paris, France: Courcier.

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in 'theory of mind'. *Trends in cognitive sciences, 8*(12), 528-533.

Lieberman, D., Tooby, J., & Cosmides, L. (2007). The architecture of human kin detection. *Nature, 445*(7129), 727-731.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences, 43*.

Liefgreen, A., Pilditch, T., & Lagnado, D. (2020). Strategies for selecting and evaluating information. *Cognitive Psychology, 123*, 101332.

Lim, J. (2012). *Welfare tradeoff ratios and emotions: Psychological foundations of human reciprocity.* Unpublished doctoral dissertation, University of California, Santa Barbara.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 986-1005.

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political economy, 115*(3), 482-493.

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., ... & Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. *PLoS one, 9*(3), e92160.

Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2020). Broken Physics: A Conjunction-Fallacy Effect in Intuitive Physical Reasoning. *Psychological Science, 31*(12), 1602-1611.

Lukaszewski, A. W., & Roney, J. R. (2010). Kind toward whom? Mate preferences for personality traits are target specific. *Evolution and human behavior, 31*(1), 29-38.

Mahr, J. B., & Csibra, G. (2018). Why do we remember? The communicative function of episodic memory. *Behavioral and brain sciences, 41*.

Marcus, G. (2008). *Kluge: The haphazard evolution of the human mind*. Houghton Mifflin Harcourt.

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).

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman and Company.

Maynard-Smith, J., & Parker, G. A. (1976). The logic of asymmetric contests. *Animal behaviour, 24*(1), 159-175.

McElreath, R., & Boyd, R. (2007). *Mathematical models of social behavior: A guide for the perplexed*. University of Chicago Press.

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.

Monroe, A. (2020). Moral elevation: Indications of functional integration with welfare trade-off calibration and estimation mechanisms. *Evolution and Human Behavior, 41*(4), 293-302.

Morris, A., MacGlashan, J., Littman, M. L., & Cushman, F. (2017). Evolution of flexibility and rigidity in retaliatory punishment. *Proceedings of the National Academy of Sciences, 114*(39), 10396-10401.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature, 434*(7031), 387-391.

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., 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.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs?. *Science, 308*(5719), 255-258.

Parker, G. A., & Maynard-Smith, J. (1990). Optimality theory in evolutionary biology. *Nature, 348*(6296), 27-33.

Pearl, J. (2000). *Causality.* Cambridge university press.

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), 209-209.

Pinker, S. (1994). *The language instinct: The new science of language and mind.* Penguin.

Pinker, S. (1997). *How the mind works.* Penguin.

Quillien, T. (2015). Population finiteness is not a concern for null hypothesis significance testing when studying human behavior. A reply to Pollet (2013). *Frontiers in neuroscience, 9*, 81.

Quillien, T. (2018). Psychological essentialism from first principles. *Evolution and Human Behavior, 39*(6), 692-699.

Quillien, T. (2020a). Evolution of conditional and unconditional commitment. *Jour-*

nal of theoretical biology, 492, 110204.

Quillien, T. (2020b). When do we think that X caused Y?. *Cognition, 205*, 104410.

Quillien, T., & German, T. (under review). *A simple definition of 'intentionally'.*

Raihani, N. J., & Barclay, P. (2016). Exploring the trade-off between quality and fairness in human partner choice. *Royal Society open science, 3*(11), 160510.

Real, L. A. (1991). Animal choice behavior and the evolution of cognitive architecture. *Science, 253*(5023), 980-986.

Sanfey, A. G. (2009). Expectations and social decision-making: biasing effects of prior knowledge on Ultimatum responses. *Mind & Society, 8*(1), 93-107.

Savage, L.J. (1954). *The Foundations of Statistics*. New York: Dover.

Schelling, T. C. (1980). *The Strategy of Conflict, 2nd edition*. Harvard university press.

Schlingloff, L., Csibra, G., & Tatone, D. (2020). Do 15-month-old infants prefer helpers? A replication of Hamlin et al.(2007). *Royal Society open science, 7*(4), 191795.

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 Behavior, 38*(6), 695-703.

Sell, A. (2005). *Regulating welfare tradeoff ratios: Three tests of an evolutionary-computational model of human anger*. Unpublished doctoral dissertation, University of California, Santa Barbara.

Sell, A., Tooby, J., & Cosmides, L. (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences, 106*(35), 15073-15078.

Sell, A., Cosmides, L., & Tooby, J. (2014). The human anger face evolved to enhance cues of strength. *Evolution and Human Behavior, 35*(5), 425-429.

Sell, A., Sznycer, D., Al-Shawaf, L., Lim, J., Krauss, A., Feldman, A., ... & Tooby, J. (2017). The grammar of anger: Mapping the computational architecture of a recalibrational emotion. *Cognition, 168*, 110-128.

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature human behaviour, 2*(10), 750-756.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics, 69*(1), 99-118.

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.

Singer, P. (2015). *The most good you can do: How effective altruism is changing ideas*

about living ethically. Yale University Press.

Shepard, R. N. (1992). The perceptual organization of colors: an adaptation to regularities of the terrestrial world?. In *The adapted Mind*, eds. Barkow, Tooby & Cosmides. Oxford University Press.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Cambridge, MA: Harvard University Press.

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science, 348*(6230), 91-94.

Sznycer, D. (2010). *Cognitive adaptations for calibrating welfare tradeoff motivations, with special reference to the emotion of shame*. Unpublished doctoral dissertation, University of California, Santa Barbara.

Sznycer, D., Tooby, J., Cosmides, L., Porat, R., Shalvi, S., & Halperin, E. (2016). 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., Al-Shawaf, L., Bereby-Meyer, Y., Curry, O. S., De Smet, D., Ermer, E., ... & Tooby, J. (2017). Cross-cultural regularities in the cognitive architecture of pride. *Proceedings of the National Academy of Sciences, 114*(8), 1874-1879.

Sznycer, D., Xygalatas, D., Agey, E., Alami, S., An, X. F., Ananyeva, K. I., ... & Tooby, J. (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., ... & Tooby, J. (2018). Invariances in the architecture of pride across small-scale societies. *Proceedings of the National Academy of Sciences, 115*(33), 8322-8327.

Sznycer, D. (2019). Forms and functions of the self-conscious emotions. *Trends in cognitive sciences, 23*(2), 143-157.

Tatone, D., Geraci, A., & Csibra, G. (2015). Giving and taking: Representational building blocks of active resource-transfer events in human infants. *Cognition, 137*, 47-62.

Tatone, D., Hernik, M., & Csibra, G. (2019). Minimal cues of possession transfer compel infants to ascribe the goal of giving. *Open Mind, 3*, 31-40.

Tesser, A., Gatewood, R., & Driver, M. (1968). Some determinants of gratitude. *Journal of personality and social psychology, 9*(3), 233.

Thomsen, L., Haugane, J., Fohn, E.K., & Born, V. (2018). Preschoolers Use the Gratefulness of Newcomers as a Cue for Their Future Altruism. *Talk presented at the 30th annual meeting of the Human Behavior and Evolution Society*, Amsterdam, Netherlands

Tooby, J., & Cosmides, L. (1990). The past explains the present: Emotional adaptations and the structure of ancestral environments. *Ethology and sociobiology, 11*(4-5), 375-424.

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In *The adapted Mind*, eds. Barkow, Tooby & Cosmides. Oxford University Press.

Tooby, J., & Cosmides, L. (1996). Friendship and the banker's paradox: Other pathways to the evolution of adaptations for altruism. In *Proceedings of the British Academy* (Vol. 88, pp. 119-144). Oxford University Press.

Tooby, J., Cosmides, L., & Barrett, H. C. (2005). Resolving the debate on innate ideas. In *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.

Trivers, R. L. (1974). Parent-offspring conflict. *American Zoologist, 14*(1), 249-264.

Von Neumann, J., & Morgenstern, O. (1953). *Theory of games and economic behavior*. Princeton university press.

Vul, E., Hanus, D., & Kanwisher, N. (2009). Attention as inference: selection is probabilistic; responses are all-or-none samples. *Journal of Experimental Psychology: General, 138*(4), 546.

Wason, P.C. (1966). Reasoning. *New horizons in psychology*, 135-151.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly journal of experimental psychology, 20*(3), 273-281.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature neuroscience, 5*(6), 598-604.

Williams, G. C. (1966). *Adaptation and natural selection: A critique of some current evolutionary thought*. Princeton university press.

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin, 24*(12), 1251-1263.

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational substrates of norms and their violations during social exchange. *Journal of Neuroscience, 33*(3), 1099-1108.

Yudkin, D. A., Prosser, A., & Crockett, M. J. (2019). Actions speak louder than outcomes in judgments of prosocial behavior. *Emotion, 19*(7), 1138.

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.