UNIVERSITÀ DEGLI STUDI
DI TRENTO
CIMeC - Center for Mind/Brain Sciences

CIMeC
Center for Mind/Brain Sciences

The Neural Computations of Trust and Reputation

A dissertation presented
by
Elsa Fouragnan
25th cycle - Brain and Cognitive Science

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Neuroeconomics

University of Trento
Trento, Italy
April 2013

Thesis advisor: Giorgio Coricelli **Elsa Fouragnan**

Abstract

Humans learn to trust new partners by evaluating the outcomes of repeated interpersonal interactions. However, available prior information concerning the reputation of these partners may alter the way in which outcomes affect learning. This thesis combines for the first time behavioral, computational, psychophysiological and neural models in a direct comparison of *interaction-based* and *prior-based* decision-to-trust mechanisms. Three studies are presented, in which participants played repeated and single trust games with anonymous counterparts. We manipulated several conditions: whether or not reputational priors were provided, the probability of reciprocation (trustworthiness) of each counterpart, and the time-horizon of the relationships.

The thesis addresses several challenges involved in understanding the complex behavior of people in social contexts, by investigating whether and how they integrate reputation into decisions to trust unfamiliar others, by designing ways to combine reputation information and observed trustworthiness into unified models, and by providing insight into information on the brain processes underlying social cognition. Numerous models, algorithms, game theoretical and neuroscientific methods are used to examine these questions. The thesis presents several new reinforcement learning (RL) models and explores how well these models explain the behavioral and neural interactions between trust and reputation.

The performance of the new models was tested using experiments of varying complexity. These experiments showed that *model-based* algorithms correlate better with behavioral and neural responses than *model-free* RL algorithms. More specifically, when no prior information was available our results were consistent with previous studies in reporting the neural detection of parametric estimates of RL models within the bilateral caudate nuclei. However, our work additionally showed that this correlation was modified when reputational priors on counterparts are provided. Indeed participants continued to rely on priors even when experience shed doubt on their accuracy. Notably, violations of trust from counterparts with high pro-social reputations elicited both stronger electrodermal responses and caudate deactiva-

tions when priors were available than when they were not. However, tolerance to such violations appeared to be mediated by priors-enhanced connectivity between the caudate nucleus and ventrolateral prefrontal cortex, which was anti-correlated with retaliation rates. Moreover, in addition to affecting learning mechanisms, violation of trust clearly influenced emotional arousal and increased subsequent recognition of partners who had betrayed trust.

Contents

Listing of figures

To Alain.

Acknowledgments

Coming from an engineering background, Neuroeconomics came as a pleasant surprise. I have been very fortunate to be part of Dr. Coricelli's team at a time when Neuroeconomics was (and is still) growing. I had the chance to work in several research centers and visit many collaborators. In particular, I spent several months in Los Angeles visiting the Economics department of the University of Southern California where I spent time with Dr. Isabelle Broca and Dr. Juan Carillo revisiting the theoretical and conceptual parts of my thesis. I thank both of them for their constructive criticisms and perspective on my work. In addition, I carried out clinical experiments with Dr. Remi Neveu in France. Whilst that work is not directly represented in this thesis, his influence, technical advice and training helped me throughout. As a "Nesquik team" member, I would also like to acknowledge the helpful inputs of Luca Polonio and Gabriele Chierchia. I learnt a great deal from their complementary characters and I am very grateful for our collaborations. I would also like to acknowledge the helpful inputs of my collaborators from the engineering department of Trento, Susanne Greiner, Paolo Avesani, Diego Sena, Andrea Mognon and Emmanuele Oliveti. Further thanks are due to Nadege Bault and Mateus Joffily for their contributions and guidance in the first and second studies. No thesis would be complete without a thesis committee. I am deeply grateful to Dr. Kerstin Preuschoff, Dr. Natalie Sebanz and Dr. Donato Grasso for taking time out of their busy schedules to generously serve on my committee and to comment on my dissertation. It goes without saying that without the training, guidance, support and friendship of Dr. Giorgio Coricelli I would not have had the confidence necessary to carry out such a project. I thank Giorgio tremendously for all these years of support concerning all aspects of my scientific career, and not least the various parts of my PhD. Last, but not least, I would also like to thank my family and my partner for the remarkable emotional support and encouragement they have provided. I would also like to thank my best friends Aurore, Audrey and Naama for staying in touch with me on a daily basis, their e-mails were definitely a welcome distraction on hard days of work!

The important thing in science is not so much to obtain new facts as to discover new ways of thinking about them.

Sir William Bragg

1

Introduction

TRUST is a critical social process that enables human cooperation to prevail in societies and organizations (Axelrod and Hamilton, 1981). However, some agents can adopt selfish and myopic behaviors that imply trust violations in order to obtain immediate beneficial outcomes (Fraser, 2011; Camerer and Weigelt, 1988). Thus relying on someone else's trustworthiness (i.e., the fulfillment of trust) is a risky decision in social exchanges because it creates vulnerabilities and exposes to the harmful consequences of trust violations. In order to reduce social uncertainty and determine whom to trust or distrust, people rely on relationship-specific history - based on previous interactions (King-Casas, Tomlin, Anen, Camerer, Quartz, and Montague, 2005; Simpson, 2007).

However, information transmitted by others (through communications or exploitation of available data, such as information acquired on the Internet) can reduce uncertainty about the outcome of social interactions (Delgado, Frank, and Phelps, 2005; Biele, Rieskamp, Krugel, and Heekeren, 2011). Indeed, reputational signals convey information about the likely behaviour of others and are potentially relevant to decide whether or not to trust - for example during online transactions. Despite the central role of both reputation information and observed trustworthiness during personal and impersonal trust-based relationships, no study

has directly confronted *interaction-based* and *reputation-based* exchanges.

Previous studies have primarily focused on unbiased, *interaction-based* trust-decisions, paying relatively little attention to the reputational component of the interaction (Krueger, McCabe, Moll, Kriegeskorte, Zahn, Strenziok, Heinecke, and Grafman, 2007; King-Casas et al., 2005; Phan, Sripada, Angstadt, and McCabe, 2010). Given the growing importance of reputation in social exchanges (e.g., the popularity of online social media) this work aims to shed light on the behavioral, computational and neural signature of trust and reputation. This thesis addresses the problem of whether, and how, reputational priors affect social decision making and learning to trust mechanisms. The work presents behavioral and computational data, *model-based* functional neuroimaging and psychophysiological measurements to characterize the role of priors in social decision making. This project proposes a new framework that takes insights from neuroscience, economics and computer science to explain neurocognitive processes involved in the decision to trust when prior information is available or not. It goes beyond previous work in its consideration of the complexities of human reasoning and making decisions with respect to reputation.

1.1 Neuroeconomics

Human decision-making and strategic thinking have been studied in a wide range of domains for decades. Cognitive scientists and psychologists use behavioral data and psychometric variables to model and explain behavioral and cognitive manifestations in controlled experimental designs. Behavioral economists attempt to understand and predict psycho-social biases that alter economic decisions in more ecological experimental settings in order to improve economic theories. Neuroscientists on the other hand, investigate the neural signature of cognitive and mental processes that guide behaviors. Conventionally, the interplay between these disciplines has been relatively restricted until the recent development of neuroscientific tools and modern technologies that have allowed the emergence a new interdisciplinary field that combines the three domains: the neuroeconomics field. Being at the crossroads between economics, psychology and neuroscience neuroeconomics aims to provide a unified theoretical framework of human decision-making. The interdisciplinary approach is vast and includes different perspectives, from neuroimaging studies to electrophysiological studies in non-human primates and incorporates various topics such as theory of choice under uncertainty, temporal discounting, framing effects, strategic choice, social decision-making, theory of mind , etc (See the book Glimcher, Fehr, Camerer, and Poldrack, 2008). The theories and models proposed in this thesis are defined by empirical evidences and by biological processes supporting human brain activities. The three studies presented in this work use game theoretical ap-

proaches, computational models and neuroscientific techniques to describe and predict human behaviours in social environments. In fact, this thesis is aligned with the "two goals of neuroeconomics" (See Montague, 2007): to understand the biological aspects of a living organism sustaining itself through time and processing information "efficiently"; and to probe the algorithms running such biological foundations (Montague and Berns, 2002; Montague, King-Casas, and Cohen, 2006; Camerer, 2007; Fehr and Camerer, 2007).

1.2 On modeling

Interactions between computer science and neuroeconomics can go both ways; algorithmic ideas can inform decision-making theories, and insights from neuroeconomics can inspire solutions of engineering problem-solving. In this work, I adopt the former view, i.e., applying the relatively sophisticated models for prediction and control from machine learning into new theories of cognitive and brain functions. This thesis aims to improve behavioral and neural theories of trust and reputation in social context by making use of theoretical and algorithmic ideas from reinforcement learning (RL). Indeed, social learning and social decision-making where appropriate actions lead to positive reinforcements (rewards) and inappropriate actions lead to negative reinforcements (punishments), can be viewed as RL problems devoted to trial-and-error and goal-directed problem-solving (Körding, 2007; Lee, 2006; Balleine, 2007). Originating from psychology, biology, and improved in computational science, RL has quickly become an important technique in neuroeconomics where it has been extensively studied in both humans and animals (Gmytrasiewicz and Doshi, 2005; Daw and Doya, 2006; Johnson, van der Meer, and Redish, 2007; O'Doherty, Hampton, and Kim, 2007; Doya, 2008).

In particular, the Temporal Difference algorithm applied to the Q-function (TD-Q learning, **see Box 1**) is thought to be instantiated in the neural reward circuitry (mimicking phasic responses of nigrostriatal dopamine neurons) (Schultz, Dayan, and Montague, 1997; Schultz, 2004), and has also been found to reflect high-cognitive behavioral and neural data (Nakahara, Itoh, Kawagoe, Takikawa, and Hikosaka, 2004; Wittmann, Daw, Seymour, and Dolan, 2008). In the family of TD models, the Prediction Error (PE) which is the discrepancy between expectation and reality, is used to update prediction values of future rewards. Additionally, by constantly minimizing PE and biasing action towards maximum future rewards, learned expectations eventually converge to optimal rewards in the long-term (McClure, Berns, and Montague, 2003). TD-RL appears to be so embedded in the literature (For an excellent review, see Dayan and Niv, 2008), that it allows testing of a wide range of hypotheses about behavioral, psychophysiological and neural data.

Box 1: Reinforcement learning

Markov Decision Process (MDP)

Formally, a MDP describes the quintuple S, A, P, R, γ, where S is the *state* space of an agent, A is the *action* space, P is the set of Markovian transition probabilities, R is the reward function such as $S \times A \times S \rightarrow, \mathbb{R}$ and $\gamma \in [0, 1]$ is the discount factor of long-term reward. The space *state* is sampled so that at each time t, the environment is in one state s_t in which the agent can choose one action a_t according to a certain policy, $\pi : S \rightarrow A$. The state then changes to s_{t+1}, and the agent receives: $r_t = R(s_t, a_t, s_{t+1})$. The agent's goal is to follow a Policy which maximises the expected long-term future reward. This quantity is called the value function and is defined for a given policy π as:

$$V^{\pi}(s) = \mathbb{E}[\sum_{t \geq 0} \gamma^t r_t | s_0 = s, \pi] \in \mathbb{R}^S$$

Given these equations, the goal is to find $\pi^* \in argmax_{\pi} V^{\pi}$. For this, the action-value function ($Q-$function) is defined. This function adds a degree of freedom on the first action that is chosen:

$$Q^{\pi}(s, a) = \mathbb{E}[\sum_{t \geq 0} \gamma^t r_t | s_0 = s, a_0 = a, \pi] \in \mathbb{R}^{S \times} A$$

The value function of the policy π^* is noted $(Q^*)V^*$. If Q^* is known then the optimal policy can be computed with a Greedy behavior as: $\pi^*(s) \in arg\, max_a Q^*(s, a)$

Bellman equations

Using the Markovian property, the value function (of a specific policy π) satisifies the Bellman evaluation equation:

$$Q^{\pi}(S_t, a_t) = \mathbb{E}_{s_{t+1}|s_t, a_t}[r_t + \gamma Q^{\pi}(s_{t+1}, \pi(s_{t+1}))]$$

To find the optimal policy, this Bellman equation can be used within a policy iteration process. A first π_0 is chosen. At interation i, the policy π_i is evaluated and the $Q-$function is excecuted. Since the policy π_{i+1} is defined as being Greedy, then:

$$Q^{\pi_i} : \pi_{i+1}(s) \in arg\, max_a Q^{\pi_i}(s, a)$$

TD-Q learning

If the model is unknown, the value function can be estimated through interactions such as:

$$Q(S_t, a_t) \leftarrow Q(s_t, a_t) + a_t(r_t + \gamma Q(s_{t+1}, aa + 1) - Q(s_t, a_t))$$

known as the Temporal Difference algorithm applied to the Q-function with a being the lerning rate.

In experiments where trust can emerge as the result of repeated interactions between partners (*interaction-based* situation), behavioral economics studies have reported that RL learning models explain existing equilibrium in game theory and the selection of particular equilibrium in case of multiple equilibria (Camerer and Weigelt, 1988; Fudenberg, Kreps, and Maskin, 1990). RL models have also been found to reflect behavioral data from people performing repeated interaction tasks (Chang, Doll, van 't Wout, Frank, and Sanfey, 2010; Bos, Dijk, and Crone, 2012). Besides capturing behavioral changes, the TD-Q learning in particular also reflects neural activities (Chapter 22 in: Glimcher et al., 2008). Investigating *interaction-based* decision to trust, recent neuroeconomics studies showed that activity in the dorsal striatum (part of the reward circuitry), shifted, over the course of interactions, from the time when trustors received trustworthy outcomes to the moment when they made their decision to trust (**see chapter 2, section 3.3**). This pattern of activity is similar to a theoretical PE signal, used to update the value of a predictive signal during learning that shift to the time of the signal itself after learning (King-Casas et al., 2005).

More formally (**see Box 1**), the original TD Q-algorithm manipulates summary representations of the Markov Decision Problem (MDP), (the value function (e.g., Q), and/or the policy π) and links received rewards into long-run expectations with the recursive Bellman Equations (Sutton and Barto, 1990, 1998; Bellman, 2003). This algorithm, named the *model-free* algorithm, can only update values along experienced state–action trajectories. Choices estimated by *model-free* systems can be compared to fast and reflexive choices based only on experience and describe very well *interaction-based* decisions (Chang et al., 2010; Bos et al., 2012). However, these algorithms have no control over the rules of MPD themselves (i.e., the probability distributions governing transitions between states and the reward function). Since these rules represent the contingencies and actual outcomes of the problem to learn, *model-free* algorithms, similar to stimuli-responses habits, cannot be adjusted by expectations, changes in goals or rules. In the present work, reputation is assumed to impact on people's beliefs about others. Therefore, *model-based* algorithms, favoring actions suggested by the agent's internal model of the environment appear more appropriate to describe prior-based decision-making where reputational priors play an important role (Doll, Simon, and Daw, 2012; O'Doherty et al., 2007).

The *model-based* versus *model-free*, distinguishing between goal-directed and habitual instrumental behaviors (Daw and Doya, 2006), will be tested in this thesis where I propose new extensions of *model-based* RL models to account for potential interactions between trust and reputation mechanisms. These models, using differential representation of the transition structure T, and the reward function R will be defined and compared to *model-free* algorithms using behavioral and neural data. This work also proposes that these two categories of models

are implemented in the brain using synergic and parallel circuits, in line with previous studies (Daw and Doya, 2006; Balleine, 2007; Tricomi, Balleine, and O'Doherty, 2009; Wunderlich, Rangel, and O'Doherty, 2009) and aims to extend our understanding of the *model-based* part of this neural architecture.

1.3 Motivations

The overall goal of this thesis is to advance theories of the dynamic interplay between reputational priors and trustworthiness beliefs from a behavioural, computational, physiological and neural point of view. The thesis is organized in three steps: (i) theoretical and computational models are proposed and then tested against behavioural data collected from a relatively large population to verify their adaptivity and validity (ii) nominated models are used to probe new behavioural data sets and their estimated parameters are entered into parametric fMRI analysis - to enrich neural theoretical models of trust and reputation interactions (iii) the new framework and theory are used in the psychological and emotional domain to validate and propose a wider range of results.

First, computationally, I propose a selection of *model-free* and *model-based* RL algorithms that can operate in a biologically plausible manner. Their properties are exposed and sensible extensions are proposed so that the new *model-based* RL systems can better explain the complexity of trust and reputation in human social context. Thus, the models' accounts of people's behaviour are made faithful, confronted directly with the presented data and theoretically with previous anomalous findings.

Second, at a brain level, the selected "best" models - in terms of behavioural-fit - are used to assess fMRI data with *model-based* parametric techniques. In this part, the thesis extends previous neuroimaging findings in trust-based decision-making with new insights on the functional interplay between the basal ganglia and prefrontal areas in human brains.

Finally, in a psycho-physiological study, the proposed theory of trust and reputation is tested with inter-individual subjective data such as emotional rating, emotional arousal recorded with electrodermal methods and eye movements.

1.4 Organisation of the thesis

The thesis first presents the relevant literature followed by the description of three experiments involving healthy humans: a behavioural/computational study, a functional neuroimaging

study, and a last experiment in which different method of psychophysiological measurement (eye-tracking and electrodermal response) are combined. All three studies make use of a game theoretical task, called Trust-Game (TG), with varying experimental parameters. In all of them *interaction-based* and *prior-based* trust decisions are directly compared (i.e., whether reputational priors were provided or not). Between each experiments, the trustworthiness level of counterparts and different relationship end points are manipulated. Each of these experiments raises theoretical issues and proposes new insights that extend existing models and theories of social decision-making in several directions.

Chapter two reviews the literature in three broad fields: (i) Economical and behavioral game theories about social learning, trust and reputation, (ii) computational work in artificial reinforcement learning and (iii) experimental evidences of behavioral and neural signatures of trust-based and prior-based social decision-making. This chapter introduces the Trust Game (TG) which is the methodological experimental task employed in the three presented experiments.

Chapter three addresses the challenges of including the effect of meaningful reputational prior in RL models, that account for external information and internal beliefs of the agent in different ways. One family of models (commonly termed *model-free* or habitual) choose actions on the only basis of experience and accumulation of reward, whereas a second (named *model-based* or goal-directed) choose actions that causally lead to reward according to people's internal representation of the situation. In this chapter, several models from these two categories are defined, adjusted and confronted with behavioral data of people playing repeated TGs. Evidence suggests that trust and reputation are intertwined in the decision-making process and that an Adaptive-Belief *model-based* RL system robustly accounts for the variability in people's interaction behavior when prior are provided.

Chapter 4 uses advanced notions from the previous study in order to explore neural activity when people play repeated TGs with stochastic endings. The *model-based* algorithm fitting behavioral data with the highest accuracy is used to estimate trial-to-trial RL parameters that are entered into fMRI analysis at time of choices and outcomes. Results suggest a functional connectivity between the prefrontal cortex, which favors actions suggested by the reputation, and the caudate nucleus that signals unexpected changes in context. Particularly exciting are the results showing that *model-based* techniques can bridge the gap between behavioral and neural results. This chapter provides a unified and normative account of both the trust-related neural responses in different areas of the brain and of how the computations these areas carry out influence learning behavior.

Chapter 5 examines the way in which varying reputational priors (i.e., from highly pro-

social to highly self-centered) influence the allocation of visual attention and electrodermal responses when participants play single TG with unfamiliar partners. This third study provides confirmatory answers about the effect of reputation in allocating trust and presents preliminary analysis on its effect on emotional reports, physiological arousal and later recognition. The most encouraging results explore the relationship between violation of social expectation and later memory-recognition and therefore should guide future experimentation and analysis in this area.

A general "law of least effort" applies to cognitive as well as physical exertion. The law asserts that if there are several ways of achieving the same goal, people will eventually gravitate to the least demanding course of action. In the economy of action, effort is a cost, and the acquisition of skill is driven by the balance of benefits and costs. Laziness is built deep into our nature.

Daniel Kahneman

2

Theoretical background

This section reviews work in neuroeconomics and in reinforcement learning (RL) related to the foundation of trust, reputation and social learning. After some material introducing key conceptual definitions, the chapter focuses specifically on previous work related to the theoretical issues relevant to the data considered in this thesis. More complete overviews of the economics and neuroeconomics foundation of trust and reputation mechanisms have been published in articles and book forms (Bohnet and Huck, 2004; Mailath, 2007; Nooteboom, 2002) and chapter in (Chapter 9 in: Rose, 2011). For reinforcement learning applied to neurosciences, I draw on other reviews and books (Bertsekas, 2012; Sutton and Barto, 1998; Kaelbling, Littman, and Moore, 1996) in what follows.

2.1 Conceptual definitions

For decades, a wide range of disciplines have demonstrated the role of trust and reputation in all human interactions. Despite their proposed importance, the definitions of trust and reputation are not fully settled. Psychology and Economics have in common some aspects of their definitions, and differ in others. Although my purpose is not to settle matters of terminology, it will be useful to clarify in which way I use the terms of trust and reputation.

2.1.1 Definitions of trust and reputation

Trust

- In Psychology - also redirected to Cognitive Neurosciences - the term "trust" characterizes both a state of mind and a social attitude. First, it refers to an individual willingness to trust, which may vary depending on the context, the personal history or an internal hormonal state (Boon and Holmes, 1991; Mayer, Davis, and Schoorman, 1995; Rotter, 1967, 1971). Second, it refers to a social attitude and social preference. Trust was defined by Luhmann (1979) as a *"process encompassing beliefs about others and their willingness to use that knowledge as the basis for action"* (Luhmann, 1979). Berg and his colleagues (Berg, Dickhaut, and McCabe, 1995) define trust as a situation in which a first person (the trustor or investor) is willing to rely on the future actions of another party (trustee). In such situation, the trustor, with or without his volition, abandons control over the actions performed by the trustee. As a consequence, the trustor is uncertain about the outcome of the other's actions; he can only develop and evaluate expectations. The uncertainty involves the risk of failure or harm if the trustee does not behave as desired.

- In Economics, "trust" refers to the situation in which a person "expects" another person to "do" something or "behave" in a certain way. Trust fits with moral hazard models due to repeated interactions and the possibility to punish "off-the-equilibrium" behaviors. In fact, the core of trust definition is that it emerges from interpersonal exchanges (Bohnet and Baytelman, 2007; Cox, 2004; Hong and Bohnet, 2004). The definition of trust is also linked with economic primitives like social preferences, beliefs about the trustworthiness of others, or risk attitudes (Coleman, 1994). Bohnet and his colleagues (2003) also identified betrayal aversion as another key feature in trusting behavior, such that it leads to the delineation of property rights and contract enforcement (Bohnet and Zeckhauser, 2003).

Reputation

- The definition of "reputation" from a Cognitive Neuroscience point of view is sparse and does not benefit from a broad range of empirical examples. Because it is still a recent topic of interest in the domain, there are differing views about the role of reputation. While studying cooperation and social dilemmas, cognitive scientists and psychologists have defined reputation as a corporate/collective image socially transmitted that pro-

vides information about people, organizations or standing (Bromley, 1993, 2000). The reputation of an agent corresponds to the aggregate of other's beliefs about that agent behaving socially. In that sense, reputation is an indirect, derive belief about other's aptitude in a social context. Reputation can be interpreted as the expectation that a trusting person holds about another person from its past actions and behaviors towards others, which include trustworthiness, cooperation, reciprocity, or norm-acceptance.

- Reputation refers in Economics to situation in which people "believe" others to act in a certain way or to "be" a certain "type" of person (Cabral, In press). It's commonly defined as *"the estimation in which a person or thing is generally held"*, as *"a favourable name or standing"* or as *"the way in which a person or thing is known or thought of"* (Deelmann and Loos, 2002). The goal of reputation mechanisms is to enable efficient transactions when cooperation is compromised by possible opportunism (moral hazard) or information asymmetry (also refers to as adverse selection[1]). Reputation has also been defined as beliefs of participants about others' strategic character (Camerer and Weigelt, 1988). The bases of the reputational mechanism are a Bayesian updating and signaling.

2.1.2 Interplay of trust and reputation

Given their definitions, trust and reputation are highly interrelated; reputation reinforces trust (and vice versa) and relates to the measurement of trustworthiness value. Indeed, one can trust another based on his reputation and its reliability. For example in online marketplaces, reputation systems have been intentionally used as trust facilitators, and both the decision-making and incentive process to avoid frauds and deceits (Kim, 2009). The trustworthiness value of an agent is computed by aggregating all reputational information that is obtained from other agents. The trustworthiness value, derived by reputation, is classified as indirect, third-party trust or transitive trust. The idea behind trust transitivity is illustrated as follow: assume Kevin trusts James, and James trusts Isabelle. When James refers Isabelle to Kevin, then Kevin derives a quota of trust in Isabelle based on James' point of view about her. This trust transitivity, defining one aspect of reputation mechanism, is illustrated in the picture below (**see picture 2.1.1**).

However, there are, at least, two main differences between trust and reputation mechanisms: (i) trust comes from direct-based interactions and the trustworthiness judgment is

[1]Adverse selection is present in situations where one person possesses information (about their type, their innate ability, the quality of what they offer, etc.) that others don't. These situations often arise in markets for experience goods.

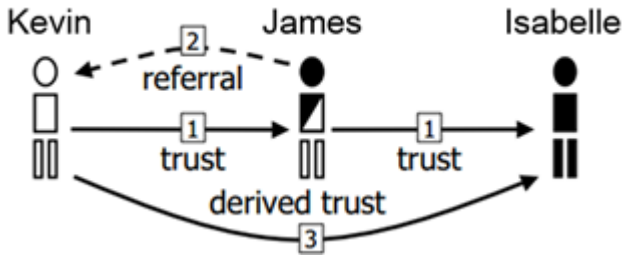


Figure 2.1.1: Trust transitivity principle. If Kevin trusts James, James trusts Isabelle and James refer Isabelle to Kevin, then Kevin derives a quota of trust in Isabelle based on James' point of view about her, in combination with his own trust in James.

a direct belief-constract, whereas reputation provides an indirect trustworthiness judgment transmitted by others, (ii) trust is based on subjective and reflexive representation of others, whereas reputational information or ratings about specific events, such as attitudes or transactions, are used as a model of trust. Given that trust and reputation are based on different sources of information, the overlap between the two is a formal certainty on the trustworthiness of others but divergence between the two information might also generate representational conflict (Deelmann and Loos, 2002; Bromley, 1993).

2.2 Reputation and trust as learning models

All definitions of trust mentioned earlier share a common assumption: trust mechanisms emerge from repeated interaction. If people interact regularly enough, they can build an *internal* image of the other that will support their willingness to trust them or not. The two next paragraphs will present key features of the economic model of trust and reputation. I will also focus on mechanisms of trust and reputation from the point of view of psychology and cognitive science, and methods to compute these models.

2.2.1 Repeated games and Folk Theorem: Bootstrap model

For economists, the repetition of a situation creates the possibility of equilibrium action profiles that couldn't exist otherwise. Within Economics, Game Theory investigates repeated and strategic thinking, and models phenomenon emerging from situations (games) where different actors interact (Deelmann and Loos, 2002; Bin Yu, 2002; Cabral, In press).

In order to illustrate trust as a social context, let's consider an interaction repeated several times between a seller and a buyer. On one hand, the seller can decide whether or not he wants to make an effort to offer a high quality product. His effort to propose a high quality product

has a cost: $effort > 0$. If he exerts effort, the product has a high quality. If he doesn't, then the chance for the product to break is high (but $effort = 0$). On another hand, the buyer surely will prefer to pay $price = 1$ for the good product that works and 0 for the one that has a high chance to break once bought. In this situation, we assume that the buyer does not know the seller so he cannot determine the seller's willingness to labor in the manufacture process.

From a theoretical point of view, the willingness to make such an effort to produce depends on the ratio between effort and price. If $effort < 1$, then it is efficient for the seller to make an effort: indeed the difference between the price paid by the buyer and the cost is still positive. However, selling a high-quality product is evidently not an equilibrium. Regardless of the beliefs of the buyers towards to seller's actions, rational economists will claim that the seller should never make any effort into selling a high quality product. In fact, when the buyer will realize that the product is not of good quality, the seller will have already been paid and its benefits will be large: $priceeffort = price$. In this case, a buyer that understands the seller strategy should never have paid for the product.

However, if the situation happens several times (game theorist would categorize this case into the "infinite repeated games with asymmetric information") then other equilibrium capture the repetition of the situation and its consequences. As an example, imagine that the buyer starts by trusting the seller and therefore pays the expected value of a good quality product, e.g. 1. If the product breaks down then the buyer has incentive to punish the seller and stop buying. In this situation, the seller should anticipate such reversal strategy and offer a good product at each time of the interaction. In terms of patience [1], one could propose that the discounted payoff from keeping his trust would be:

$$\left(\frac{1 - effort}{1 - price}\right) \tag{2.1}$$

If he decides at any time to not invest effort on the product he will get a 1 in the short-term but 0 in the long-term. Therefore trust in this scenario is an equilibrium if (**Equation 2.1**) is >= 1 or if $price > effort$.

One of the central tenets of the repeated game theory, the folk theorem, states that if players are patient enough, then any rationally possible set of individual payoffs can be sustained as the Nash equilibrium of a repeated game (Cabral, In press). The folk theorem can be interpreted as a model of trust: If players are patient and the future matters a lot to them, then mutual trust

[1] We can relate a player's discount factor δ to her patience. How much more does this person value a dollar today, at time t, than a dollar received $\tau > 0$ periods later? The relative value of the later dollar to the earlier is $\delta^{t+\tau}/\delta^{\tau}$. As $\delta \leftarrow 1$, so as her discount factor increases she values the later amount more and more nearly as much as the earlier payment. A person is more patient the less she minds waiting for something valuable rather than receiving it immediately. So we interpret higher discount factors as higher levels of patience.

can exist in equilibrium. The basic idea is very simple and intuitive: there is a trade-off between a myopic short-term gain from defeating the other and a long term loss from destroying trust and reputation.

2.2.2 Reputation as a Bayesian model

The fundamental feature of trust emergence in economics models was discussed earlier: the bootstrap mechanism explicitly states that trust emerges from repeated interactions when deviations of trust can be punished. Now let's imagine that, the seller could adopt either a good or a bad profile in each stage of the repeated interaction: (i) he could decide to be trustworthy and sell a high quality product with a high probability to function: a_H or (ii) he could decide to be untrustworthy and create a low quality product that has a low probability to function: a_L, where $0 < a_L < a_H < 1$. On the other hand, the buyer is risk neutral and thus will always offer a price that is equal to his willingness to pay for that product. We can formalize this situation with extra parameters: if a product works, it is worth 1 unit and if it doesn't, it is worth 0. If δ is the buyer's belief that the seller is trustworthy, then buyer will be willing to offer a price defined as:

$$price = \delta * a_H + (1 - \delta) * a_L \tag{2.2}$$

With this equation, the price will increase as a function of δ and pertains to the buyer's belief that the seller is trustworthy. In this particular situation, this could be qualified as the reputation of the seller, which will determine how much the buyer is willing to invest in the seller's product. Let δ_0 be the prior belief that the seller is trustworthy, T the number of times that product was of high quality and U when the product was of low quality. Thus the following equation captures the update of the buyer's belief:

$$\frac{\delta_0 a_H^S (1 - a_H)^F}{\delta_0 a_H^S (1 - a_H)^F + (1 - \delta_0) a_L^S (1 - a_L)^F} \tag{2.3}$$

Where the prior belief δ_0 determines the seller's history (the sum of T and U) and the reputation of the seller is given by δ. The value of δ is a function of T and U. If T increases, then the reputation of the seller is better and if U increases, the reputation will decrease.

Consequently, if the reputation of the seller increases, the price will also increase (in the same way, the reputation decreases after proposing low quality products). Finally, this equation captures another phenomenon: if the seller is trustworthy then δ moves closer to 1 and the price moves closer to a_H. If he is untrustworthy, the price converges to a_L, where $a_L < a_H$.

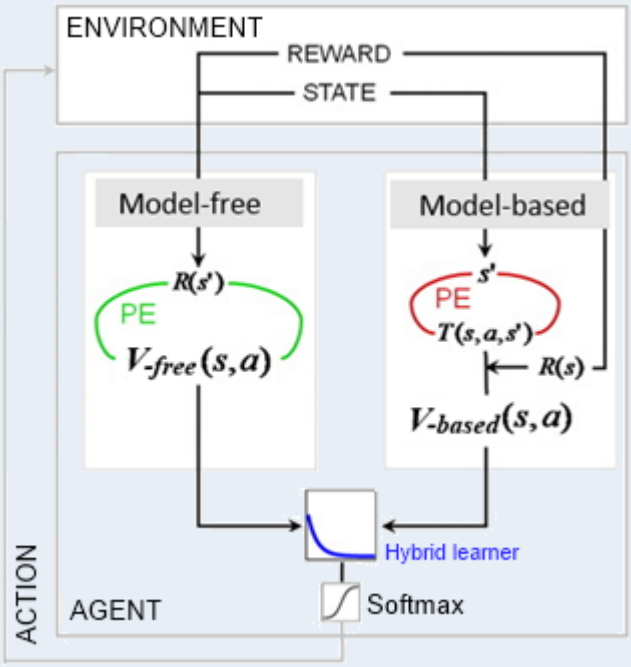
2.2.3 Trust and reputation: Reinforcement Learning problems

During the last decade, studies started to apply modeling in conjunction with empirical and behavior data to better probe social decision-making. In their experiment, Hampton and colleagues used computational approaches to investigate the ability of mentalizing about others while playing interactive social game (Hampton, Bossaerts, and O'Doherty, 2008). Furthermore, Apesteguia, Huck, and Oechssler (2007) reported that when watching others playing a game, people tend to imitate the strategy of the winner of the game (Apesteguia et al., 2007), behaviors that can be captured by computation modeling. Using social games, an extensive literature on the questions of optimal strategies and decision-making proposes to use reinforcement-learning theories to provide insights in how people solve such optimization problems (Lee, 2006; Balleine, 2007; Körding, 2007).

Model-free and model-based RL algorithms

Two candidate families of RL models have figured prominently in recent neuroeconomics theories of learning with evidences from the human and the animal literature. The first comes from the *"law of effect"*, which states that agents *"habitually repeat actions that have been reinforced in the past"* (Thorndike, 1911). The second comes from the notion of a *"cognitive map"*, an internal representation of the environment that agents utilize to plan goal-directed sequences of actions (Tolman, 1948). These two types of learning algorithms are named "*model-free*" and "*model-based*" systems (**see box 2**). As presented in Chapter 1, TD-Q algorithms are called *model-free* because they only learn an error-driven update function (they only learn from previous experience) that closely resemble the activity of dopaminergic neurons. In contrast, *model-based* RL algorithms (dynamic programming) are based on a more flexible and richer internal model of the learning environment (**see box 2**).

Box 2: *Model-free* and *Model-based* algorithms



Theoretical model for data analysis using both a *model-free* learner and a *model-based* learner. *Model-free* computes a PE using direct experienced reward and state from the environment to update state-action values. The *Model-based* learner learns a model of the state space T(s,a,s') by means of a state-PE, which is then used to update the state transition matrix. Action values are computed by maximizing over the expected value at each state. Then, a hybrid learner computes a combined action value as an exponentially weighted sum of the action values for the *model-free* and *Model-based* learner. The combined action value is then submitted to softmax action selection. (Figure inspired by Gläscher et al. (2010).)

1. **Model-free computation**, illustrated in the left part of the picture uses experience – state and reward - to learn directly the characteristic of a problem – state and action values -. Given a decision policy, a state has a value which is defined as the future utility that is expected to increase starting from that state. One advantage of this computation is that it doesn't need an internal representation of the world to make a decision, it is a rather guileless and automatic way to learn without any *a-priori* of the world. On the contrary, this computation has the disadvantage that (1) a lot of trial-and-error experience is required to correctly predict the state's values that lead to future consequences, (2) the model cannot adapt quickly to changes in the environment and (3) incorrect information from the environment would be also combined with previous estimates without the possibility to discard them.

2. **Model-based computation**, illustrated in the right part of the picture refers to an internal representation of the problem "a mental map" that has been learned based on observation that does not concern the current problem itself. This internal representation includes knowledge of the features of the task, (formally, the probabilities of transitions from one state to another and different immediate outcomes), which can also be conceived as introspective thinking in human psychology. This is a statistically efficient way to use experience, since each piece of information from the environment can be memorized in an internal representation that can be re-use. Another advantage of this model is that it adapt quickly to changes (action selection changes rapidly in the transition contingencies and the utilities of the outcomes). The inconvenience of the model is that it needs a support for estimating the world *a-priori*.

A wide range of behavioral and neural studies proposes that the brain employs both *model-free* and *model-based* decision-making processes, with each process dominating the other in certain situations (For a review, see Dickinson and Balleine, 2002) Therefore, different neural substrates are suggested to underpin each process (Balleine, 2005; Daw, Niv, and Dayan, 2005).

2.2.4 Neural correlates of Model-based vs. Model-free implementations

Animal and human studies have reported that agents use *model-free* and *model-based* form of learning processes under different circumstances (Dickinson, 1985; Philiastides, Biele, Vavatzanidis, Kazzer, and Heekeren, 2010). Besides, the executive control of the behavior can alternate between the two systems (Wimmer, Daw, and Shohamy, 2012; Wunderlich, Symmonds, Bossaerts, and Dolan, 2011; Beierholm, Anen, Quartz, and Bossaerts, 2011) and specific brain lesions might impair one system while preserving the second, suggesting that the brain contained separate, competing systems for *model-based* and *model-free* RL - a habit RL system in which choices are selected based on previous experience and a goal-directed system, more adaptive, that implements a cognitive map (Dickinson and Balleine, 2002; Doll et al., 2012).

Recent electrophysiological studies indicate that the dorsolateral striatum mediates *state — reward* habits and thus play an important role in the *model-free* implementation theory (Yin, Knowlton, and Balleine, 2004). However, other electrophysiological studies have shown that *model-based* behaviors also require the integrity of the dorsomedial striatum, suggesting that the striatum plays an important role both in *model-based* and *model-free* implementations (Tricomi et al., 2009; Zweifel, Parker, Lobb, Rainwater, Wall, Fadok, Darvas, Kim, Mizumori, Paladini, Phillips, and Palmiter, 2009; Tsai, Zhang, Adamantidis, Stuber, Bonci, Lecea, and Deisseroth, 2009). The dichotomy between *model-based* and *model-free* is still a recent topic in fMRI studies. However, converging results indicate *model-based* value signals in ventromedial, ventrolateral prefrontal cortices (vmPFC, vlPFC) and adjacent orbitofrontal cortex (OFC) which plays an important role in goal-directed sequence of actions (Valentin, Dickinson, and O'Doherty, 2007; Wunderlich et al., 2009; Jones, Somerville, Li, Ruberry, Libby, Glover, Voss, Ballon, and Casey, 2011). Thus, two brains areas, (i) the prefrontal cortex which is commonly associated with executive functions – higher order cognitive processes that manage, control and regulate other brain activities (Ernst and Paulus, 2005), and (ii) the limbic system which has generally been associated with emotion regulation and more recently with social processes (Bush, Luu, and Posner, 2000; Rilling, Gutman, Zeh, Pagnoni, Berns, and Kilts, 2002), seem to signals *model-free* and *model-based* information respectively.

2.2.5 First Hypotheses and Motivations

Whether trust and reputation are distinct constructs or represents a trust–reputation continuum is a debated topic in the literature. While some studies hold these two concepts as separate, some others studies integrate the two in an unified model (Bromberg-Martin, Mat-

sumoto, Hong, and Hikosaka, 2010; Hiroshi Abe, 2011). However no studies so far has tried to employ the RL framework described in the previous section to test whether trust and reputation can be employed as variable in *model-free* and *model-based* RL algorithms. Taking insights from dynamic programming my attempt is to propose a *model-based* RL algorithm that integrates reputation priors into the decision-making problem to trust (or not) someone during social interaction (**See figure 2.2.1**).

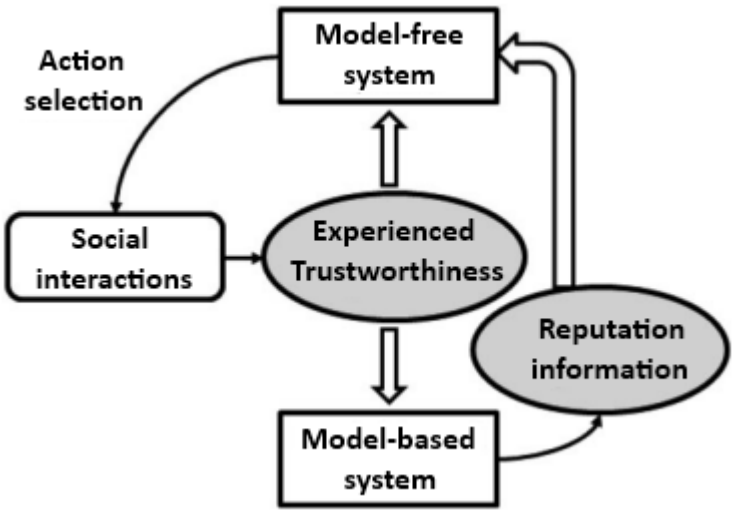


Figure 2.2.1: Trust and reputation interactions in RL algorithms. Figured inspired by Sutton when describing the dichotomy between *model-based* and *model-free* systems (Sutton and Barto, 1998). The social interactive environment provides the participant a direct measure of trustworthiness (states and rewards), whereas the *model-based* system furnishes the agent with expectation about the other driven by its reputation cues. The *model-free* system applies the same policy to select an action from both real trustworthiness and belief based on reputation.

Indeed, a *model-based* algorithm involve explicit or implicit secondary structure (such as counterfactual signals, rules stored in memory or cues from the environment) where information about rewards that are not actually received can be inferred or observed. Thus, a pure *model-free* RL would be blind to reputation information and only update decision-values to trust according to observed trustworthiness. However, decision to trust with reputation priors go beyond *model-free* RL and allow to directly manipulate the key features of *model-based* RL (for the formal definition see **box 2** and **chapter 3**), that is, the computation of decision-values using prior expectation that influence the reward function and a sequential transition model of an action's consequences. The *model-based* learns the transition and possible reward of the interaction and uses this model to generate predictions about future reward.

2.3 Behavioral and neurobiological evidences

Politicians and economists have begun to pay attention to the importance of trust on economic flows and progress when they realized that *"much of the economic backwardness in the world could be explained by the lack of mutual confidence"* (Arrow, 1972). As we have seen earlier, theoretical economic models explain that trust can be a best response in situation when an agent needs to build social capital (Lewis and Weigert, 1985). These models were proposed to reconcile empirical evidences from the assumption of pure rationality. Indeed, humans genuinely trust unrelated strangers and are cooperative even in a double-blind situation, (e.g. when they will never interact with a person again or when they have no possibility to cultivate their own reputation) (Berg et al., 1995; McCabe, Houser, Ryan, Smith, and Trouard, 2001). Therefore decision to trust and reciprocate trust are thought not to be directed at furthering one's own interests but rather are guided by other-regarding preferences (Fehr and Camerer, 2007; Falk and Fischbacher, 2006; Van Lange, 1999).

However reciprocity and trustworthiness have been found to be a behavior dependent on individual differences in social value orientation (SVO) which characterizes inter-individual tendency to value the outcome of others (Van Lange, 1999; Lahno, 1995).

2.3.1 Assessing trust

Trust is a nebulous construct. It is difficult to isolate, quantify, and characterize as an experimental cognitive mechanism. Nevertheless, to assess trust in experimental set ups, it is necessary to operationalize it, although this effort may be limited.

The Trust Game

Camerer and Weigelt (1988) developed one of the earliest methods of measuring trust and trustworthiness in a laboratory setting through an economic game which was as a streamlined version of the investment or Becker–DeGroot–Marschak method (BDM)[1] investment game pioneered by Berg, Dickhaut and McCabe (Berg et al., 1995). This task is called the Trust Game, has been employed and modified for many years and is one of the widest spread method

[1]The Becker–DeGroot–Marschak method (BDM), named after Gordon M. Becker, Morris H. DeGroot and Jacob Marschak for the 1964, Behavioral Science paper, "Measuring Utility by a Single-Response Sequential Method" is an incentive-compatible procedure used in experimental economics to measure willingness to pay (WTP)

to operationalize trust in laboratories, by quantifying both trustworthiness and the willingness to trust. The mechanics of the game are presented in **figure 2.3.1**.

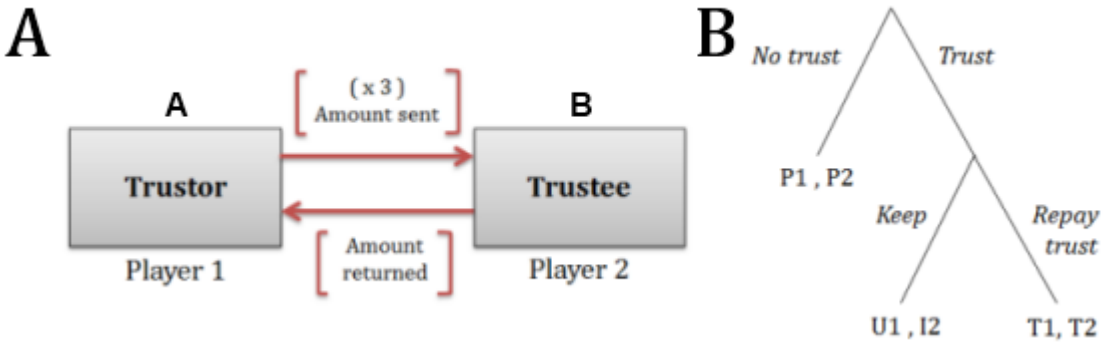


Figure 2.3.1: The "trustor" is the first player (A), and the "trustee" is the second player (B). The two players' payoffs are given as (P, T, U and T), where P_1 is the trustor's payoff and P_2 is the trustee's payoff at the beginning of a game. Player A can choose between Trust and No Trust. If A chooses No Trust then the game ends, without B making a decision, and the payoffs are unchanged ($P_1 = P_2$). If A chooses to trust B, he gives his entire payoff to B, who then chooses between keeping everything, with payoffs ($U_1 < I_2$), and reciprocate in which case the payoffs are equally divided ($T_1 = T_2$, and $T_1 > P_1$).

In another version of the TG, the first stage of the game consists of the trustor's decision to invest any portion x of his endowment to the trustee. If A decides to invest an x amount of his money, x will be subtracted from his endowment until the next move. The experimenter will multiply the money invested, x, by a certain factor $f > 1$, before sending it to the trustee. In stage two, B has the choice to either keep the endowment or reciprocate trust and give any portion y of the money he received back to the trustor.

The TG is made of two sequential stages of economic exchanges in which no one is contractually committed to enforce agreements. In behavioral economics in particular, many participants come to the laboratory and receive an initial endowment that is known by all participants. Then participants are anonymously paired with each other and assigned to either the role of trustor (A) or trustee (B).

The amount of money sent by the trustor and the trustee are said to capture trust and trustworthiness respectively. The money sent by the trustor captures the willingness to bet that the trustee will reciprocate, which is a risky move (at a cost to them). Moreover, the decision to trust is linked with an expectation that the consequences of the decision will be positive in terms of the trustor's investment. Particularly, if the second mover repays trust, the trustor would be better off than if the trustor doesn't invest to begin with. This definition of trust as a

behavior was proposed by Coleman (1994; 1995) and can be precisely observed using game theoretical one-shot games performed anonymously (Coleman, 1994; Berg et al., 1995).

Cross-national data

The Trust Game (TG) has come to dominate the field as a way to operationalize trust and trustworthiness. Outside of laboratories, studies investigating on trust use questionnaires on large population. One example is the American General Social Survey which is an annual survey on trust since 1972, and the World Values survey which probes multicultural differences in trust. Both surveys capture trust using the following question: *"Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?"* The possible responses are either *"Most people can be trusted"* or *"Can't be too careful"* [1].

These questions and corresponding responses have been found to be relatively problematic. Indeed, a person who is not willing to take small risks even with potential gain could agree to both choices, since his opinion about people is independent of his general attitude towards risk (Miller and Mitamura, 2003). To rule out this problem, some authors have proposed one dimensional question that separate trust and distrust, such as: *"Do you think that people can be trusted?"* with ratings from *"Absolutely"* (rated 7) to *"not at all"* (rated 1) as possible answers.

Other tools have been developed to measure trust determinants (Bohnet and Huck, 2004; Bohnet and Baytelman, 2007; Eckel and Wilson, 2011; Jamison, Karlan, and Schechter, 2006; Houser, Schunk, and Winter, 2010), and trust in international data sets. These tools allow two main advances: (1) to probe the effect of organizations on trust (Bohnet and Huck, 2004) and, (2) to perform cross analysis on both organizational and political comparisons of trust (Naef and Schupp, 2009; Porta, Lopez-de Silane, Shleifer, and Vishny, 1996; Houser et al., 2010).

2.3.2 Behavioral evidences

Berg and his colleagues (1995) found a significant willingness to trust and reciprocate trust among participants that were engaged in one-shot and anonymous TGs – a result that deviates considerably from Nash predictions and simple maximization of people's own monetary

[1] The question is very close to asking people about their behavioral inclinations "can't be too careful" and it seems quite likely that when people answer this question, they consult either their own experiences or behaviors in the past or introspect how they would behave in situations involving a social risk. Therefore, it seems likely that the answer to the GSS and WVS question is not only shaped by people's beliefs about others' trustworthiness, but also by their own preferences towards taking social risks.

payoffs (Berg et al., 1995). Since (1995), studies using the TG have repeatedly reported trust-related behaviors and reciprocity across gender, age, culture, geography and category socio-professional and it has shown that majority of trustors in the game send more than half of their initial endowment and the majority of trustees reciprocated with more than the amount sent by the trustor. (For a meta-analysis of the trust-game see Johnson and Mislin, 2010).

In repeated TGs, studies report an increase of trust over time when the trustee has proven to be trustworthy. The theory is rather intuitive: people are more likely to trust someone that was proven reliable than someone who has the potential to betray them. In "TG terms", the best predictor of whether a person will place trust in their partner in a given round is whether or not this partner previously reciprocated trust. However, if someone betrays us then we stop trusting that person. This process, also named a tit-for-tat strategy, has been reported to be one equilibrium in repeated social games (Boyd, 1989) [1].

2.3.3 Neural correlates of trust and trustworthiness

Having proved its usefulness and explicability, the Trust Game has been recently employed in fMRI settings investigating the neural bases of trust (King-Casas et al., 2005), cooperation (Decety, Jackson, Sommerville, Chaminade, and Meltzoff, 2004) and reciprocity (Phan et al., 2010; Krueger et al., 2007). Most of these fMRI studies have reported that BOLD signal in the medial prefrontal cortex (mPFC) was higher when participants were interacting with a human counterpart compare to a computer (McCabe et al., 2001; Rilling et al., 2002), and when participants decided to trust compare to when they decided not to trust (McCabe et al., 2001; Krueger et al., 2007; Phan et al., 2010; Delgado et al., 2005; King-Casas et al., 2005). Hence, it has been proposed that decisions to trust are associated with high activation in the prefrontal cortex.

The literature has also reported that the dorsal striatum (caudate nucleus) plays a role in signaling the magnitude of an observed social reward (positive and negative) (Knutson, Adams, Fong, and Hommer, 2001). Rilling et al (2002) showed that activation in the striatum correlates with positive or aversive interactions in a Prisoners Dilemma Game (Rilling et al., 2002; Rilling, Sanfey, Aronson, Nystrom, and Cohen, 2004) and predicts behavioral reciprocity for mutual gain (Rilling, King-Casas, and Sanfey, 2008). In another series of studies, activation of the dorsal striatum also differentiated between positive and negative outcomes in repeated TG (Delgado et al., 2005; Stanley, Sokol-Hessner, Fareri, Perino, Delgado, Banaji, and Phelps, 2012). Additionally, Krueger and colleagues (2007) reported that the calculation of rewards is associated with activation in the dorsal striatum (Krueger et al., 2007). Therefore, decision

[1] In financial aspect, learning to trust allows the trustor to maximize his net payoff.

to trust or not appears to be the consequences of observations leading to activations in the caudate nucleus.

In other studies where participants were receiving fair and unfair treatments (during Ultimatum Games), authors found activation in the insular cortex (Sanfey, Rilling, Aronson, Nystrom, and Cohen, 2003; Tabibnia, Satpute, and Lieberman, 2008). These finding are supported by recent studies which report that untrustworthiness was associated with enhaced activations the insula, anterior cingulate cortex and caudate (Chang and Sanfey, 2009; Stanley et al., 2012). It has therefore been suggested that these brains areas play a role in signaling personal norm violations (King-Casas et al., 2005) and signaling positive and negative aspects of social exchanges.

However, while this pattern of activity is in line with a wide range of studies, King Casas and his colleagues (2005) were the first to find that these regions were computing social prediction error, theoretically responsible for triggering social learning on the bases of trial-to-trial reward-learning (Balleine, 2005; Daw et al., 2005). Because this study was pivot in the field, the next paragraphs is dedicated to their experimental paradigm and their results.

Authors of this important study asked 48 pairs of participants to play the TG repeatedly against each other for 25 rounds while their neuronal activities were recorded by two simultaneous fMRI acquisitions (hyperscanning techniques as described in **figure 2.3.2**). The two participants were assigned randomly to play as trustor or trustee and then keep their role for the entire scanning session.

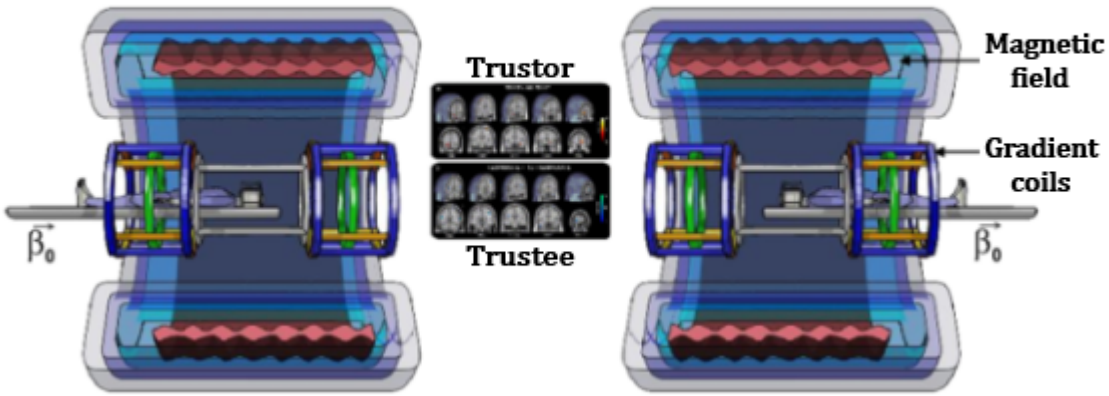


Figure 2.3.2: Hypersacnning of the Trust Game. Hyperscanning is a new fMRI acquisition method where two or more participants interact with each other while being scanned in different MRI scanners.

First, in order to categorize player moves, the authors specified three categories: (1) benevolent reciprocity, where the trustor gave a generous amount despite a decrease in repayment by the trustee, (2) malevolent reciprocity, where the trustor invested a lower amount after an increase in repayment by the trustee and finally (3) neutral reciprocity, where there was no change in investment and repayment moves. The corresponding analysis revealed differential BOLD activities in brain regions of trustors that encountered a benevolent or malevolent trustee reciprocity behavior compared to neutral reciprocity behavior. These results suggest that there is a unique response to deviations in the expected behavior of one's partner, particularly in the caudate nucleus, the inferior frontal sulcus, the superior frontal sulcus, the thalamus, and the inferior/superior colliculli.

Second, in order to compare the neural responses in dorsal striatum when trustees showed benevolent or malevolent behaviors, the authors performed region-of-interest (ROI) analyses and revealed that the BOLD signal in the striatal voxels correlated with the trustee's subsequent decision to be trustworthy. Indeed, the trust expressed by the trustor was found to predict the future changes in trustworthiness expressed by the trustee (King-Casas et al., 2005).

Last and most importantly, cross correlation analyses were performed in the trustor's dorsal striatum in order to find if participants learnt to trust (or not) their partners as the game progressed. Consequently, changes in the striatum were examined across early, middle, and late rounds (see **figure 2.3.3**) using cross-brain and within-brain correlational analysis. The authors found that early caudate's response to trustworthy behavior, at time of outcome, shifted to the time of the decision, relating for the first time to the shift of reward prediction errors common to *model-free* RL systems, but in the context of a social exchange.

This pivotal study gave rise to the currently held view that the reward circuitry computes efficiently reward-harvesting problems in a way similar to *model-free* RL algorithms. This fMRI study was the first of a series confirming that the dorsal striatum is processing outcomes information (de Quervain, Fischbacher, Treyer, Schellhammer, Schnyder, Buck, and Fehr, 2004; Kable and Glimcher, 2007) in order to learn and adapt choices through trial and errors (Schönberg, Daw, Joel, and O'Doherty, 2007).

Finally, other studies using fMRI in clinical patients also found that, along with other subcortical areas, the amygdala and midbrain areas that are involved in the processing of fear, menace, risk and social betrayal, play an important role in the decision to trust. The amygdala has been found to be involved in avoiding social presence and agoraphobia (Adolphs, Tranel, and Damasio, 1998) and its activity also increases when seeing untrustworthy faces (Engell, Haxby, and Todorov, 2007). A decrease in amygdala activation has also been found to be re-

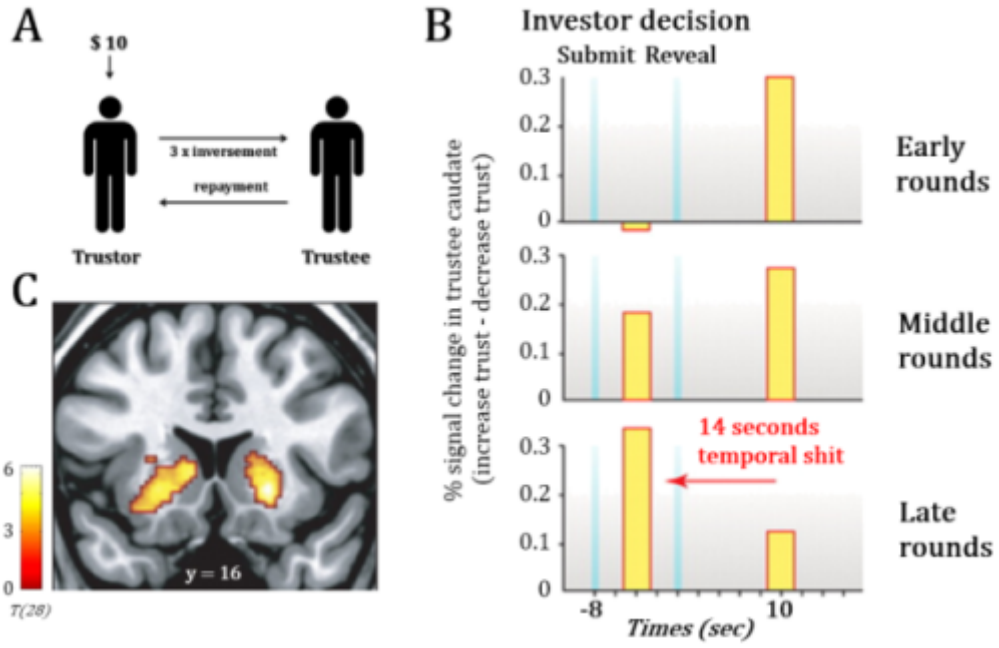


Figure 2.3.3: Activation of the striatum when reciprocal behavior is revealed. A. Trust game played simultaneously between two players while recording their brain activities. B. Schematic of authors reports: they segregated BOLD responses in response to the revelation of the decision to trust (time = 0 s). Hemodynamic amplitudes increase when first trust is revealed. As the game progressed (middle and bottom), the peak of this differentiated response underwent a temporal transfer from a time after the revelation of the investor's decision (t = 14 s; an anticipatory signal). C. ROI were performed in the head caudate of the trustee and revealed activity in the striatum.

sponsible for hyper-sociability (Evans, Wright, Wedig, Gold, Pollack, and Rauch, 2008). Additional clinical investigations have linked bilateral damage in the amygdala with the inability to judge the trustworthiness of peoples' faces. People suffering from these lesions showed a propensity to judge others as either good or bad, with no gradient in between (Öhman, 2002).

2.4 The human capacity to track priors and social cues

Trial and error is not the only method for learning predictability. Indeed, experiments in psychology have proven that trustworthiness can be quickly inferred from information not directly related to the interaction, and can also affect behavior in the Trust Game. Thus, decision makers assimilate advice or relevant social cues that discount feedback from interpersonal history as information to guide their decisions (Biele et al., 2011). Humans also infer beliefs about others from prior information or social cues. Perceptions of moral character are sufficient to modulate the dorsal striatum during the decision and the outcome phases of a trust game (Delgado et al., 2005). In this study, participants were presented with descriptions of life events showing exemplary, neutral, or suspicious moral character of their hypothetical coun-

terparts in the repeated TR. Results showed differential activations in the striatum between positive and negative feedback, as well as between no-trust and trust decisions with the "neutral" partner. No differential activity was observed for the "good" partner despite experimental manipulations to ensure that neutral and good partners responded in the same way. These findings suggest that prior information can diminish the reliance on brain structures, such as the dorsal striatum, that are important for behavioral adaptations to feedback information. These brain structures are also of interest to our experiment as the subjects may feel the need to adjust their behaviors following inconsistent outcomes. Therefore, if prior information diminishes the behavioral adaptation to this feedback, such an effect might be modulated by a diminished reliance on brain structures involved in behavioral adaptation.

Not only humans use information about moral character (Delgado et al., 2005), but also group membership and ethnicity (Stanley et al., 2012) to judge trustworthiness. Even more subtly, facial expressions processed outside of conscious awareness (Morris, Frith, Perrett, Rowland, Young, Calder, and Dolan, 1996), smiling (Centorrino, Djemai, Hopfensitz, Milinski, and Seabright, 2011), body posture (Ekman, 1992) or attractiveness could be enough information to infer a person's trustworthiness before any direct exchange. Other studies report that competence judgments about an individual can be made within 100 ms (Willis and Todorov, 2006) and affective judgments about an individual can be made as quickly as 160 ms (Pizzagalli, Lehmann, Hendrick, Regard, Pascual-Marqui, and Davidson, 2002). The presence of relevant cues is thought to correlate with activation of the dorsolateral prefrontal cortex (Delgado et al., 2005; Li, Delgado, and Phelps, 2011), a region responsible for emotion regulation and cognitive control. Priors unrelated to direct interpersonal evidence affect initial belief and have a sustaining effect on a person's learning mechanism. A recent neuroeconomics study (Chang et al., 2010) investigating how initial judgment can affect trust also reported that the implicit assessment of facial traits can predict the amount of monetary risk that a trustor is willing to take in a TG. Indeed, in one-shot TG, pre-ratings of trustee's facial trustworthiness predicted the financial amount trustors were willing to invest in them.

Additionally, another study also found that following the advice of third-party alters the BOLD responses in the reward system with an increase of activity in the septal area and the left caudate (HC) when the advice is followed (Biele et al., 2011).

Finally, relying on someone else's trustworthiness involves uncertainty about the potential risk of betrayal (Pavlou and Gefen, 2005). A recent meta-analysis of studies of decision-making under uncertainty has reported the vlPFC, vmPFC and OFC cortices to be the primary areas for processing uncertainty (Krain, Wilson, Arbuckle, Castellanos, and Milham, 2006). In one study, it has been found that the activation in these areas increases in magnitude with higher levels of uncertainty (Huettel, Song, and McCarthy, 2005), whereas other studies

reported that the OFC cortex can distinguish among different uncertainty levels (Hsu, Bhatt, Adolphs, Tranel, and Camerer, 2005; Gonzalez, Dana, Koshino, and Just, 2005). Thus, higher degree of uncertainty, for example at the beginning of an exchange with an anonymous partner, is hypothesized to be associated with a higher activation in the prefrontal and orbitofrontal cortices.

The creator of the universe works in mysterious ways. But he uses a base ten counting system and likes round numbers.

Scott Adams

3

Behavioral and computational evidences

3.1 Motivation

Previous studies in Neuroeconomics have started to investigate the cognitive and neural processes underpinning human cooperation and trust. Several factors, including available prior information and personal history have been proven to affect people's propensity to trust in subsequent interactions. However, very few studies have been focusing on the interaction between these two factors during direct exchanges. Our first aim in this chapter is to present how reputational priors interact with experienced trustworthiness in a repeated Trust Game (RTG). We made the following hypotheses: (1) available prior information on the reputation of others and direct evidence reliably influence decision to trust, (2) these two types of information happen independently and synergistically when reputational prior are provided, and (3) reinforcement learning (RL) models can be used to test whether trustworthiness are based on reputational priors, probability of reciprocation or their interaction. This first study aims at providing a novel quantitative framework that theorizes trust and reputation during social exchange.

3.1.1 Reinforcement learning framework

Because of its simplicity, not only the RL paradigm is a robust way to investigate how humans learn from feedback, but this framework also allows us to probe how social signals from different sources affect learning during repeated interactions. In this first study, our goal is to use RL algorithms to investigate adaptive social learning during RTGs and to propose new insights into the processes underpinning adaptive social decision making. Despite the importance of trust and reputation, no studies to date, have attempted to confront the two situations (decision to trust with or without reputation information) and relatively few studies have analyzed social learning with an RL perspective (Behrens, Hunt, Woolrich, and Rushworth, 2008; King-Casas et al., 2005; Chang et al., 2010).

Nevertheless, using non-social learning tasks, some recent studies have applied computational models to investigate whether advice provided by a third party would impact learning mechanisms (Biele, Rieskamp, and Gonzalez, 2009; Biele et al., 2011). In these experimental settings, participants received instructions, also named advice, from either other participants or the experimenter, about the best decision to make. In these presented experiments, authors have shown that social prior impacts probabilistic-learning mechanisms and leads to learning bias, (e.g. correct information guides participants toward good decision-making while incorrect information damages learning). In a first paper, Biele et al. found evidence supporting the notion that an "Outcome-bonus" RL model best fits the behavioral data compare to any other RL models. This model proposes that outcomes consistent with an advice will be gratified a learning "bonus" while outcomes inconsistent with advice would be disregarded (Biele et al., 2011). In another paper, Doll et al. (2009) suggest that the model explaining more of the variance in behavioral data initializes the value of decisions that were recommended with a greater starting value. Thus, the decision value would be reduced in case of inconsistent outcomes while amplified in case of consistent outcomes (Doll et al., 2009). Taken together, these studies propose that information transmitted from a third party influences both initial judgment and the way people would learn from feedback. According to these findings, reputational priors consistent with outcomes would be weighted higher and lead to a more important update while reputational prior inconsistent with outcomes would be weighted less. However, no study to date has examined how reputation information impacts learning in an interactive social scenario.

3.1.2 Hypothesis

In the present study, we investigated how explicit reputational priors interact with observed level of trustworthiness (i.e. the probability of reciprocation of a given counterpart) in a simplified version of the RTG. We hypothesized that reputational priors influence initial social risk-taking in an economical exchange (Delgado et al., 2005; Dominic S Fareri, Luke J. Chang, and Mauricio R. Delgado, 2012; Stanley et al., 2012). Secondly, we predicted that observations from direct-based interactions also impact on behavior (Axelrod and Hamilton, 1981; King-Casas et al., 2005). Lastly, and most importantly, we expected that these two signals, reputational priors and observed reciprocity, would interact in the learning process.

We thus analyzed behavioural data of participants playing the repeated version of the TG at two time points: at time of choice and at time of outcome and used RL models to assess three different hypotheses - (1) Initial decision to trust would be biased by priors, - (2) the trustworthiness belief about counterparts would be updated according to direct feedback, and - (3) participants would have a dynamic representation of the environment. We used different RL models dealing differently with the initial values of decision to trust and the way outcomes are updated. We expected that eventually the two processes could coexist: initial values can be overridden by the participant's real observations (i.e. whether or not the counterpart was trustworthy or not during exchanges). For example, among all models that we tested, the "Outcome-Bonus" model suggests that initial reputational priors would influence the way outcomes are updated during interactions. This model also implies that initial expectations bias learning in its direction (a reputation for being cooperative adds weight to a reciprocative move and would disregard violation of trust). Another RL model, the Prior-Expectation model suggests that information about other's reputation would be used as an initial belief that would be progressively dominated by experience. Inside the RL paradigm, we proposed a new hybrid model which first predicts that reputation information is used to initialize trustworthiness belief and then is dynamically updated on the basis of observed trustworthiness: the Adaptive-Belief model. Thus, the reputational prior influences both expectations and learning mechanisms. By operationalizing the potential cognitive mechanisms involved in decision to trust and reputation via several RL models, this first study aim was to increase our understanding of how reputation priors and learning to trust mechanism interact in social economic exchanges.

3.2 Experimental Design and Task

3.2.1 Task

Participants played a simplified version of the RTG in the role of the investor as described in the introduction. We employed a 2 by 2 within-participants experimental design in which the counterpart's type of reputation (unknown or known) was crossed with the counterpart's likelihood for reciprocity (high or low). Each counterpart represented one of the four experimental conditions. Level of reputation was assessed with the Social Valuation Orientation (SVO) task: an independent pro-social questionnaire (See description below).

Participants played a simplified version of the RTG in the role of the investor (or trustor) as described in the introduction. Each trial lasted 11 seconds and began with a short fixation cross (0.5 second) followed by a picture of the counterpart (3.5 seconds). Participants then decided to trust or not the counterpart by pressing "keep" or "share". After submitting their investment, counterpart's decision was computed and the outcome of the game was displayed in the screen and the trial (as well as the TG stage) ended. If the participant did not submit an offer in time (in a windows of 2 seconds), a message was displayed, telling that they have lost the trial (**See figure 3.2.1**).

We defined the counterpart's likelihood for reciprocity as a high (80%) or a low (20%) probability of reciprocation. The money first allocated to the participant was 1 euro. If invested by the participant, this first endowment was multiplied by a factor of 3. If an offer was reciprocated, counterpart always reciprocated 50% of the total multiplied amount sent by the trustor. When trust was not replaced, the counterpart kept all the multiplied amount of money. Participants played 10 randomly interspersed rounds with each counterpart (400 trials total).

The pictures for the human counterparts were collected from the FERET program (Phillips, Moon, Rizvi, and Rauss, 2000). The words "trust" or "trustworthy" were never mentioned . For each participant the condition in which the prior was provided and the one in which it was not were randomly assigned to an experienced trustworthiness condition (high vs. low). The four cells were balanced across participants, $\eta^2(3) = 0.57, p = 0.93$. All stimuli were presented on a laptop via Presentation software (Neurobs inc.).

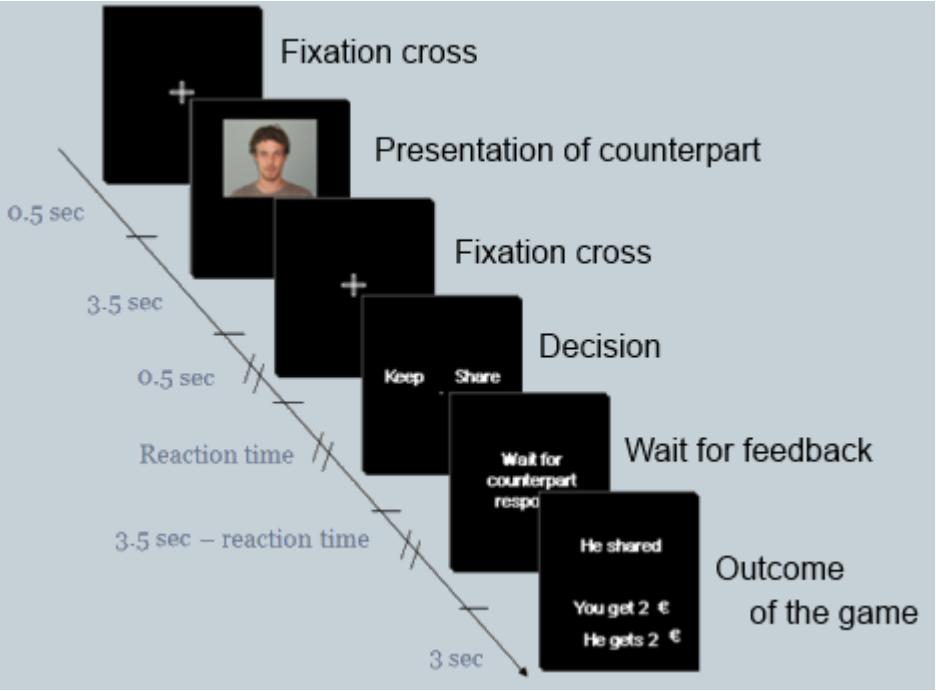


Figure 3.2.1: Partners were introduced by seeing the picture of their counterpart for 3.5 sec. Then, they made sequential decisions as first mover (trustor) presented in a binary way: "Share" or "Keep". They can choose to keep and quit the game with its initial endowment or can choose to share and continue the game. The choice of the second mover is then computed while a message is displayed on the screen " Wait for counterpart response" and the feedback of the game is displayed on the screen for both players (3 sec).

3.2.2 Participants

A total of fifty-four healthy graduate students were recruited from the University of Lyon 1 and 2, France (29 females). They all had previously joined the recruitment system on a voluntary basis and were recruited via an online annoncement in which they were screened to exclude any with a history of psychiatric or neurological disorders. Two participants were excluded from the data after indicating that they did not understand the experiment. We report results from fifty-two participants (29 girls, Mean age = 19.56 SD $\pm$ 2.17). These participants gave written informed consent for the project which was approved by the French National Ethical Committee.

3.2.3 Social Value Orientation

We instilled reputational priors by providing a cue (triangle and circle) during the presentation of counterparts in the task. Participants were told that these priors indicated the counterparts' scores on the Social Value Orientation (SVO) measure (**See figure 3.2.2**).

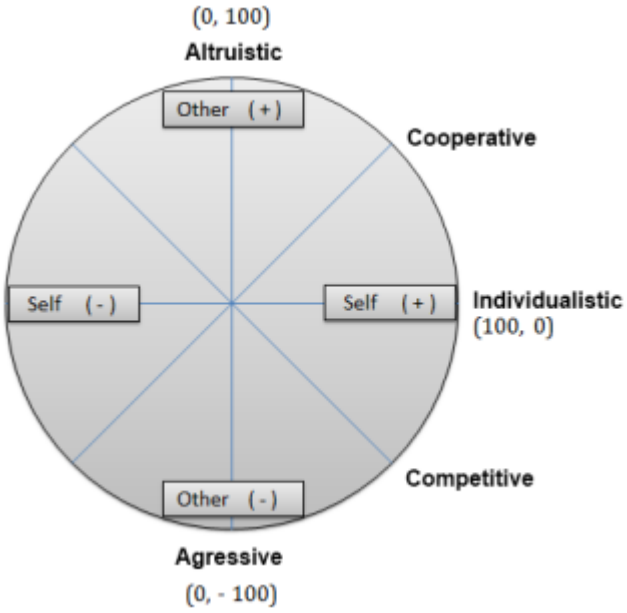


Figure 3.2.2: The value orientation ring. Individuals with vectors lying between 67.5 and 112.5 degrees are altruistic, with vectors between 22.5 and 67.5 degrees are cooperative, with vectors between -22.5 and 22.5 are individualistic, with vectors lying between -67.5 and -22.5 are competitive and with vectors between -112.5 and -67.5 are aggressive.

The SVO task is a well-established questionnaire in social psychology (Van Lange, 1999) that distinguishes between different types of SVO (among all are the cooperative and individualistic type). The main difference between each category is the extent to which one cares about own payoffs and that of the other in social dilemma situation. In our study we used responses of fictive counterparts matching the profiles of cooperative and individualistic persons which was reflected by two probabilistic way of reciprocating during the game. Cooperative counterparts show a probability of reciprocation = 80% whereas individualistic counterparts show a probability of reciprocation = 20%.

3.2.4 Experimental Setup

Each experimental session lasted 400 trials. Participants sat in the behavioral room and played on a computer. They were told that all the responses of the second players were recorded and replayed dynamically as the game progressed. They were also told that their own gain would depend on their decisions and on the decisions of the others. In each trial, participants chose to share or keep their allocated money (1 euro). At the end of the experiment, 10 random trials were extracted from the data and given to the participant [1].

[1] This manipulation ensures that each trials are equally motivating for the participants.

3.2.5 Statistical analysis

All analyses were performed with the statistical software package Stata, (College Station, TX, Release 9/SE). A test for normality was performed to verify that data was normally distributed (Shapiro-Wilk).

3.3 RL Modeling

3.3.1 Action selection

To test whether our behavioural data reflected RL processes, we examined responses of different computational RL models. All models made use of a reward prediction error (RPE) values to update beliefs associated with each choice (share or keep) that determine the probability to make the decision with the maximum value (Schultz et al., 1997; Egelman, Person, and Montague, 1998; Holroyd and Coles, 2002; Schultz, 2004; Montague and Berns, 2002). For example, a negative outcome of the game generated a negative RPE which was used to decrease the value of the chosen decision option (e.g., to keep or share with a given participant), making the model less likely to opt for the same decision on the following trial.

Precisely, the probability P of deciding to share on trial t is the logit transform of the difference in the value V_t associated with each decision, computed with a Softmax policy function (Sutton and Barto, 1981; Montague, Hyman, and Cohen, 2004):

$$P_{(share)t} = \frac{exp^{(\frac{V(share)_t}{\tau})}}{exp^{(\frac{V(share)_t}{\tau})} + exp^{(\frac{V(keep)_t}{\tau})}} \tag{3.1}$$

where the action value $V_{(share)_t}$ corresponds to the expected reward of sharing and τ is called a *exploration* parameter. For high *exploration* values ($\tau \to \infty$) all actions have nearly the same probability (and the model would select an action randomly) and for low *exploration* values (i.e., $\tau \to 0^+$) the probability of the action with the highest value tends to 1.

3.3.2 Model-free and model-based RL

The probabilities of reciprocation had to be learned through the experiment. Therefore we decided to run several model-based RL algorithms that give more flexibility to the reputation information or to the observed trustworthiness. We compared the performance of several computational RL models when fitting estimates of the models with the participants' behavioral

responses. We chose the following six models and their legitimate combinations according to various hypotheses on the role of priors, learning rates, outcome update and their interactions for two main reasons: (1) All these RL models have hypothesized neurobiological signatures (Montague et al., 2004), and thus allow experimental predictions suitable for neuroscience studies; (2) All parametric changes are appropriate economics variable that account for human variability in uncertain situation.

Baseline model: *model-free* RL

The first model implemented is the classical TD learning (For more details, see **Chapter 1, Box 1**). In short, if we tease apart how rewards and decision policies are learned, then the value of the decision is equivalent to the immediate reward plus the discount value of future rewards (that are the direct consequences of the probability function defined in **section 3.3.1**). One way to compute this sum of discounted future rewards is to sample the future rewards in a recursive way (Bellman equation)[1]. Thus, the value of a current decision is defined as:

$$V_t = R_t + \lambda . V_{t+1} \tag{3.2}$$

where R_t is the current reward, V_{t+1} is the next future decision value (that I also named state's value in the **Chapters 1 and 2**) and λ is the discount parameter. Thus, decisions values can be implemented directly with an update rule such as:

$$V_{t+1} = V_t + \alpha[R_t + \lambda . V_{t+1} - V_t] \tag{3.3}$$

or more simply, if δ_t is the prediction error such that: $\delta_t = R_t + \lambda . V_{t+1} - V_t$, then:

$$V_{t+1} = V_t + \alpha . \delta_t \tag{3.4}$$

which corresponds to the classical TD rule where α is the learning rate.

Note that, in the equation, there is no temporal discounting between the decision and the feedback. In this task, it would not make sense to discount values of action from the time of the decision until the time of the feedback, because we fixed the time-window between the two events at 3.5 seconds and because the outcomes unambiguously resulted from the preceding decision. However, the model discounts weights from previous and future trials rather than

[1]Knowing a priori all states and decisions that will be made in the future requires a precise model of state flow (including future decisions) and tracing all possible future branches is computationally intensive, time consuming and does not match with the complexity of real environments.

discounting the action value before receiving the feedback, in line with other studies (Barraclough, Conroy, and Lee, 2004; Cohen, 2006; Cohen and Ranganath, 2007).

Gain Loss model

The "Gain and losses" model is based on evidence supporting the Prospect Theory which represents an important achievement in behavioral Economics and studies of decision-making under uncertainty (Tversky and Kahneman, 1979). These authors demonstrated how people underweight feedbacks that are merely probable as compare with certain feedbacks. Evidences supporting this theory have also demonstrated that decision values are differentially assigned to gains and losses as compare with final rewards. In short, the authors found that people prefer avoiding losses as compared to obtaining gains of the same magnitude. Thus, we propose to test a second RL model that is an extension of the TD learning but that also weights differently subjective utility representing losses and gains. Specifically, our second model differentially updated gains and losses via separate learning rates (Doll et al., 2009; Chang et al., 2010). This model computes a predicted value for the next trial for each decision (share or keep) based on the experienced outcome such as:

$$V_t + 1 = V_t + a^G.\delta_t^G + a^L.\delta_t^L \tag{3.5}$$

where a^G is the amount that a positive outcome is weighted and a^L is the amount that a negative outcome is weighted in the update ($0 < a < 1$). This allows people to learn from losses in a different way than gains. We assumed this particular feature to be of particularly importance in a social task. Indeed, evidences show that sensitivity to social gains and losses motivate participants differently than in individual contexts (Bault, Coricelli, and Rustichini, 2008). Other findings reveal that participants are more willing to switch their behaviors after violation of trust (Bohnet and Zeckhauser, 2003) rather than unexpected benevolent move. For this model in particular, we chose to set the initial value V_0 for all conditions to the average amount sent by the participants on the first trial of the game (mean = 1.52 euros).

Prior expectation model

The following models where inspired by the "Novelty-bonus" model introduced by Dayan in 2002 (Kakade and Dayan, 2002). These models, the simple Prior-Expectation model and the Prior-Decay model, both initialize the initial values of the classical TD learning model according to the nature of the prior information. In this task, the Prior-Expectation model assumes that, when reputation priors are provided, participants have "optimistic" or "pessimistic" ex-

pectations, at the beginning of the game due to the presence of a positive (P^+) or negative Prior (P^-), respectively (Biele et al., 2009, 2011; Wittmann et al., 2008). Thus, the values of initial choices when playing with a Cooperative or Individualistic counterpart in the prior condition were formally defined as:

$$V^{Prior+}_{(C,o)} = g^{P+}.\mu\theta_{Prior+}.N \tag{3.6}$$

$$V^{Prior-}_{(C,o)} = g^{P-}.\mu\theta_{Prior-}.N \tag{3.7}$$

where g^{P+} and g^{P-} are equal to 1 when playing with a counterpart with a positive or negative prior, respectively; and 0 otherwise. θ_{P+} and θ_{P-} are free parameters capturing the optimistic or pessimistic impact of the priors expectation, μ is the expected payoff from choosing randomly among all options, which serves as a normalization constant and N is the number of trials experienced in the learning condition, which is a scaling factor, allowing for the comparison between an expected decision value and the outcome of the decision. On the other hand, in the no prior condition, only one parameter weighted the initial expected value of choices, $V^{NP}_{(C,o)}$. The Prior-Expectation model predicts that the reputational prior will only influence the initial expectations and will have no effect on the update function.

Prior-Decay model

An alternative hypothesis, the Prior-Decay model is that reputational prior influences initial trustworthiness belief, but then becomes less important with growing evidence when playing with a counterpart. This is a different test of the Prior-Expectation hypothesis because it predicts that the reputational prior will influence the update from the beginning of the interactions and not just influence the initial value itself. This model influences the decision values as a function of time by ρ and is computed such that:

$$V_t + 1 = V_t + a.\delta_t + e^{-\rho.V_{(C,o)_t}} \tag{3.8}$$

where $0 < \rho < 1$. This model shares some assumptions with the Outcome-Bonus for the early period of the interactions, but exponentially reduces the influence of the positive and negative prior expectation over time. The Prior-Decay model tests the Prior-Expectation hypothesis seen earlier and predicts that reputational priors will provide a first estimate of an counterpart's level of trustworthiness, but will eventually be overwhelm by observed trustworthiness, in line with previous studies (Chang et al., 2010).

Outcome-Bonus model

The "Outcome-Bonus" Model, also named the Confirmation-Bias model (Friedrich, 1993) was first proposed to account for prior information that do not necessarily influence initial decision values but rather impact on the way decision values are updated. This model is based on previous research that has examined the effect of explicit information provided by a third party (such as advice, information of moral character or facial expression) on learning tasks (Biele et al., 2009, 2011; Wittmann et al., 2008; Doll et al., 2009; Chang et al., 2010; Doll, Hutchison, and Frank, 2011). These studies have proposed models that give a higher weight to feedback that is consistent with the given information, and a lower weight to feedback that is inconsistent with the transmitted information. To examine this hypothesis we tested a model that is similar to Biele and colleagues (2011) implementing a stochastic prediction error δ computed across conditions such as:

$$\delta_t^P = r_t - V_{C_t} + g^P.\mu.\theta_v \tag{3.9}$$

where g^P is a dummy function that takes the value 1 if option C corresponds to the one suggested by a Prior and 0 otherwise, θ_v is a free parameter capturing the level of reputational prior influence and μ is the expected payoff from choosing randomly among all options that serves as a normalization constant (in our case $\mu = 1$).

Adaptive-Belief model

Our final hypothesis is based on a new model that takes propositions from the Prior-Expectation and the Outcome-Bonus models. The Adaptive-Belief model also uses reputational prior as a bonus in the update function, but rather than being a constant value based on the initial belief like the Outcome-Bonus and Prior-Decay models, it dynamically changes over time based on the experienced probability of reciprocation. This implies that the Adaptive-Belief model can induce the trustworthiness belief $P_{(t,TW)}$ from a reputation signal for each counterpart and will interpret this information as a bonus or deduction in the update function. We formally define this new model as:

$$T_{(0,TW^+)} = g^{Prior+}.\mu\theta_{Prior+}.N \tag{3.10}$$

$$T_{(0,TW^-)} = g^{Prior-}.\mu\theta_{Prior-}.N \tag{3.11}$$

$$T_{(t+1,TW)} = T_{(t,TW)} + \beta.[v_t - T_{(t,TW)}] \tag{3.12}$$

$$V_t + 1 = V_t + \alpha.\delta_t + \theta[T_{(t+1,TW+)} - T_{(t+1,TW-)}] \tag{3.13}$$

where β is the learning rate of the trustworthiness belief and θ is a free parameter representing how much the learning bonus influences the decision value update. P is initialized as in the Prior-Expectation model: $g^P.\mu\theta_P.N$ and depends on the nature of the reputational prior, if available. In **equation 3.12**, we set the value of v as 1 if the counterpart is trustworthy or 0 if the counterpart violates trust.

Thus, the Adaptive-Belief model dynamically assesses the level of trustworthiness of each counterpart by using the reputational prior and rewarding a bonus proportional to the experienced level of reciprocation or alternatively by subtracting a proportional value if the counterpart violates trust. This model is different from the Outcome-Bonus model because violation of trust by a counterpart who has a bad reputation implies a smaller deduction to the update function. Indeed, the Adaptive-Belief allows prior reputation to influence the update function similarly than the Outcome-Bonus model but it also allows feedback to modify the trustworthiness beliefs. Since this new model captures the experienced level of reciprocity of each counterpart, it can also be used to assess participant's sensitivity to trust.

3.4 Model evaluation

To estimate the fitness of all our models, we run a cross-validation technique which involves using half of the original data set as the validation data, and the remaining observations as the training data. This method considerably reduces over-fitting issues as well as it provides a convenient way to compare models of different complexity. During the first (also named training) phase, each model was compared to the participant's actual behavioral by minimizing the sum of the squared error (SSE). Free parameters were computed for the entire group of participants. Then, the comparison between models was performed by extracting the most parsimonious model. Finally, we tested the ability of each model to predict the behavioral data in the remaining observation data (unused during the classification). This method allows to controls for the number of free parameters included in the model when fitting to behavioral data (Hampton et al., 2008).

The TG investment by the participant was multiplied by a factor of 3 and added to the starting endowment: 1 euro. Then they were divided by 2 because they were split equally between both parties (when the counterpart reciprocated). The participant's outcomes thus corresponded to 2, 1 or 0 euro, received at the end of each trial, which allows the possibility for

participants to still have a positive monetary amount after at the end of a trial even when there is negative prediction error (i.e. the counterpart was not trustworthy but participant kept). After each trial, a prediction error (PE) was calculated. For example, in the classical RL model, the PE was the difference between the weighted outcome received (0, 1 or 2 monetary units) and the specific weight for the chosen target (e.g., $\delta_1 = 0 - V(share)_t$) in the case in which the counterpart decided to keep after a decision to trust made by the participant. Free parameters were estimated during the training step by minimizing the SSE between the observed data of our participants and the predicted data from the different models using *"fminsearch"* (Coleman and Li, 1996), a multivariate unconstrained nonlinear optimization algorithm implemented in Matlab (Mathworks, Cambridge, MA).

$$\sum (r(s)_t - V(s)_t)^2 \tag{3.14}$$

where V is a decision (or a state value) value s at time t, and updated at time $t+1$ with the functions specified above. The parameters were estimated for the entire group of participants because each RTG was only composed of 10 trials.

This manipulation implies that whereas the different models were fit to every participant's individual data, the error generated when estimating the free parameters was pooled across all participants. Indeed, this technique, when employed for individual parameter estimates, has proven to not be stable for a small number of trials and when the collinearity between parameters is large (which is clearly the case for some of our models)(Daw and Doya, 2006). As a result, we found that a hierarchy of model works for individual parameter that fits each participant (we report the more stable group fits). In addition, in order to reduce start location in local minima, we used a grid search algorithm for the initial values of these parameters.

All models were compared to the first model described, the baseline RL model (**Equation 3.3**), using both the Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC), two statistical metrics of model-fit that differentially take into account the complexity of the model (by penalizing differently the number of additional estimated parameters). AIC is formally defined as:

$$AIC = 2k + n[ln(\frac{2.\pi.RSS}{n} + 1)] \tag{3.15}$$

where k is the number of free parameters, n is the number of observations, and RSS is the residual sum of squares (Akaike, 1974). On the other hand, BIC is formally defined as:

$$BIC = k.ln(n) + n[ln(\frac{2.\pi.RSS}{n} + 1)] \tag{3.16}$$

Finally, we tested the model with the cross-validation procedure by randomly extracting half of the samples from the observation data set that was retained as the validation samples for testing the model and the remaining observations were used for training. We extracted the parameters that were estimated during the training phase and used them to minimize the SSE resulting from the difference between the behavioral data and the model predictions (and since no parameters were estimated during this approach, the result of the model-fitness was not the result of additional free parameters).

3.5 Results

3.5.1 Behavioral Results

Overall, we found a main effect of the probability of reciprocation on the participants' decision to trust. Participants trusted more counterparts who reciprocated 80% of the time (mean = 72.16, se $\pm$ 5.18) as compared to counterparts who only reciprocated 20% of the time (mean = 31.36, se $\pm$ 3.81) using a repeated measures ANOVA $F(1, 51) = 119.13$, $p < 0.001$, $\eta^2 = 0.66$.

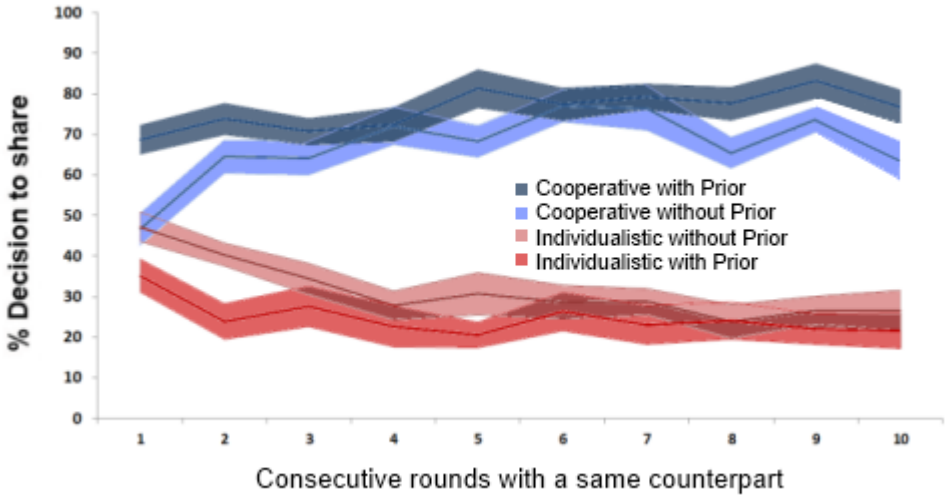


Figure 3.5.1: Overall learning in the iterated Trust Game. Note: this figure shows the mean investment amount across subjects for each trial of each condition.

There was a significant reputation by reciprocity interaction $F(2, 102) = 5.25$, $p = 0.006$, $\eta^2 = 0.08$, such that participants trust more reputed cooperative counterparts that reciprocated 80% of the time (mean = 78.12, se $\pm$ 4.29) and less reputed individualistic counterparts that reciprocated 20% of the time (mean = 25.64, se $\pm$ 4.02).

Finally, this observed interaction was not entirely driven by a strong effect of reputation at the beginning of the experiment. Indeed, we observed a significant interaction between reputation and probability of reciprocation on the last trial of the experiment $F(2, 102) = 4.31$, $p = 0.02$, $\eta^2 = 0.06$, suggesting that the effect persists throughout all 10 trials (**See figure 3.5.1**). These data were log transformed to account for negative skew in the data.

We found that participants made more money when playing with cooperative counterparts compared with individualistic counterparts. In **figure 3.5.2**, we have plotted for each of our participants the average payoff of all interactions with trustees with cooperative reputation as blue cross, and the average payoff of all interactions with trustees with individualistic reputations as green open circle. We graph the joint earning of the 52 pairs in a large outer triangle with point (0,4), (1,1), and (2,2) which indicates the set of feasible earning pairs.

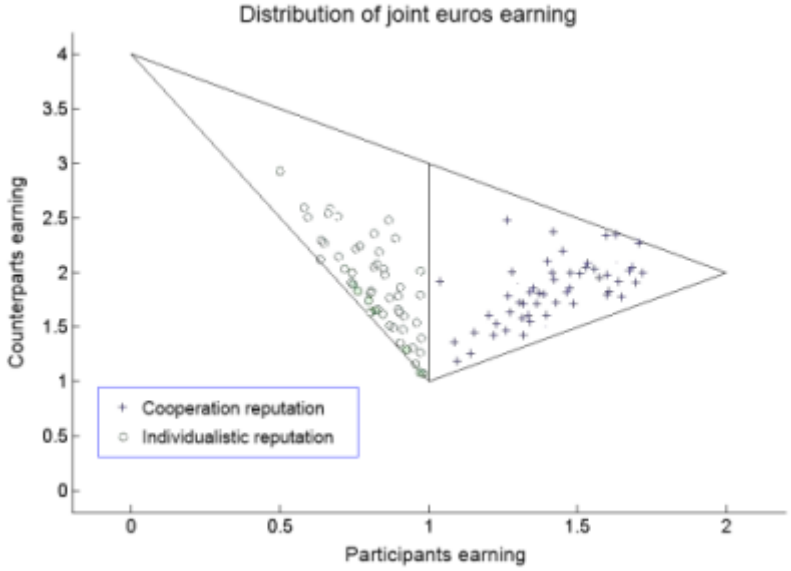


Figure 3.5.2: Representation of trustor's and trustee's earnings during the trust game. Each participants average earning playing with a certain counterpart is represented by a cross or circle. Blue crosses are average payoff of all interactions with trustees with cooperative reputations and green open circles are the average payoff of all interactions with trustees with individualistic reputations. The outer triangle shows the set of possible earning pairs.

The triangle (0,4), (1,3), and (1,1) indicates the earning pairs with non-positive net returns to participants, while the triangle (1,1), (1,3), and (2,2) indicates the non-negative net returns to participants. The data suggest that reputation increases the non-negative net returns to participants and this increase is statistically significant (p-value of paired t test = 4.41^{-09}). Coupled with counterparts with cooperative reputation they gain more (mean = 1.468, se $\pm$ 0.145) than with counterparts with individualistic reputation (mean = 0.818, se $\pm$ = 0.132)

3.5.2 RL results

All the models that we used explained more variability than the baseline TD RL model (BIC = 7632; RSS = 2265) resulting in best model-fits results. In details, both the Outcome-Bonus model (BIC = 7366; RSS = 2134) and the Prior-Expectation model (BIC = 7069; RSS = 2201) explained better the data compare to the baseline model (**See figure 3.5.3**) . Finally, the Prior-Decay model (BIC = 6432; RSS = 2231) and the Adaptive-Belief model (BIC = 6489; RSS = 2133) provided very good fit of all the RL models tested. The one exception was the GL Initialization model alone, which did not appear to fit the data any better than the baseline model (BIC = 7678; RSS = 2311). However, once combined with the other models, the GL model added some statistical power to some models. Indeed, the best fit of all model is the combination of the GL and the Adaptive-Belief model (BIC = 6318; RSS = 2144).

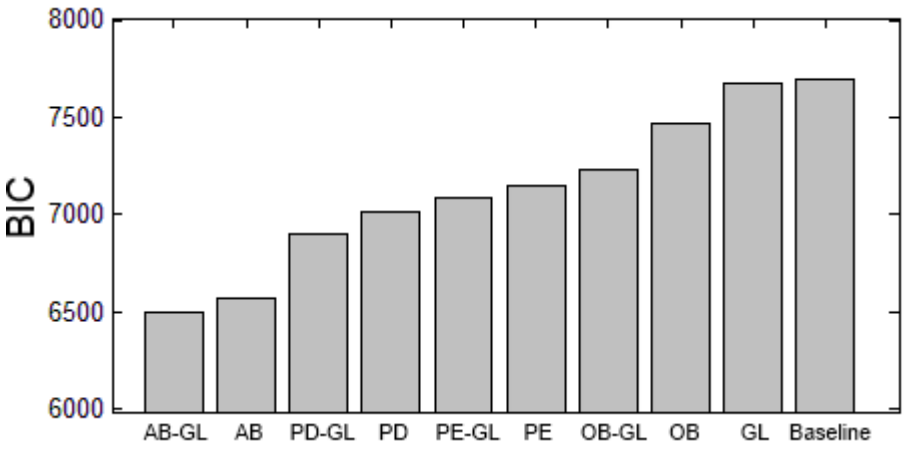


Figure 3.5.3: This graph depicts the Baysesian Informetion Criterion for RL models in which every model was fit to half of the trials and parameters were estimated for the entire group. Note: AB, Adaptive-Belief model; GL: Gain Losses model; PD: Prior-Decay model; PE: Prior-Expectation model; OB: Outcome-Bonus model.

These first results suggest that reputation priors do not just merely influence first impressions represented by initial decision values in RL models but rather seem to impact the way outcomes from interactions are interpreted. The combined GL and Adaptive-Belief model gathering information from the reputational prior and from the outcomes of the game in a dynamic fashion appeared to be the best account of the behavioral data. This model predicts that the reputational prior would impact both initial value and the update function differentially for positive and negative reputation and also accounts for the feedback to update the expectation. Nevertheless, it is important to note that the combination of GL and Adaptive-Belief models also results in a model that contains more free parameters compare to the others. While the AIC and BIC metrics are standard techniques of penalizing for additional free parameters in

Performance of the models estimated with a cross-validation procedure		
	BIC	RSS
Baseline (*model-free*) TD learning	7721	2332
Gain Losses Model	7708	2187
Outcome Bonus Model	7488	2436
Prior-Expectation Model	7133	2345
Prior-Decay Model	7004	2216
Adaptive-Belief Learning	6566	2431
Outcome Bonus Model with separate update for Gain Losses	7249	2299
Prior-Expectation Model with separate update for Gain Losses	7117	2317
Prior-Decay Model with separate update for Gain Losses	6819	2198
Adaptive-Belief Learning with separate update for Gain Losses	6499	2219

Table 3.5.1: Learning best model Averaged best-fitting parameter estimates (across subjects) SE

theoretical model, we decided to investigate further our results by using the first half of the data to predict the remaining trials. Since no further parameters are estimated during this test, then the results are unbiased by free parameters.

Using this rotation estimation, we found a similar hierarchy of results (**See table 3.5.1**). The GL model alone gave results that are closer to baseline TD learning algorithm (BIC = 7708; RSS = 2187). The Outcome-Bonus model (BIC = 7488; RSS = 2436) and the Prior-Expectation model (BIC = 7133; RSS = 2345) both fit the data better than the TD learning algorithm and both the Adaptive-Belief model and the Prior-Decay model (BIC = 7004; RSS = 2216) exhibits the best fit. These results provide additional evidence that reputation signal about trustworthiness influence particularly the update function, and that outcomes in turn can update judgment.

3.5.3 Further results on best RL model

We tested all RL models in two different ways. In a first step, we explored whether the participants' behavioral data fitted with RL mechanisms. This was achieved by providing to the model the decision's history and experimented reinforcements (from each participant, we used their real actions during RTGs as inputs to the models) and comparing empirical behavioral data to the RPE and values of the two decisions (share and keep) that the model generated for each trial and each participant. In order to perform these analyses, we mathematically estimated free parameters for each participant using a maximum likelihood minimization function provided by Matlab *fminsearch* (MathWorks, Natick, MA) (Barraclough

et al., 2004; Luce, 2005; Cohen, Heller, and Ranganath, 2005). This algorithm uses the non-linear, unconstrained Nelder–Mead simplex method to find appropriate free values of the RL model parameters that maximize the two probabilities to share and keep across the whole experiment.

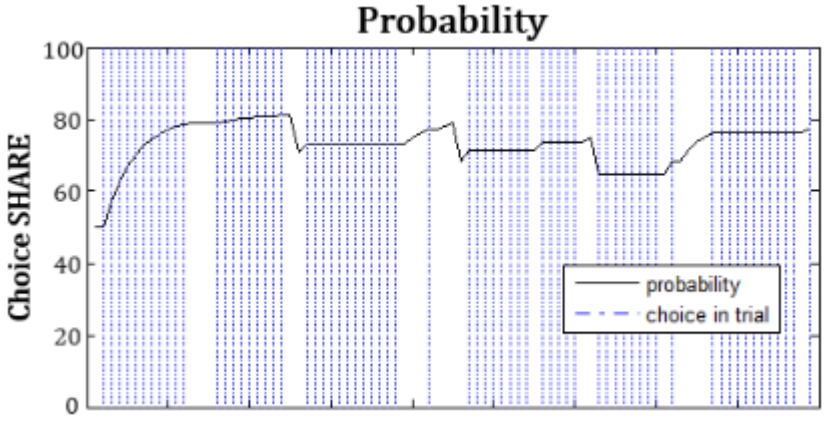


Figure 3.5.4: Example of the simulation of one RL model based on real outcomes for the TG

In a second step, we used the resulting best model of the first part to take the role of the participants and performed the RTG against each counterpart. This was done to infer the behavior of the model and examine its prediction errors and decision values as it played the RTG (for an example of one simulation, **See figure 3.5.4**). For these analyses, we used the set a_t to 0.4 and λ to 1 for both positive and negative feedback. To explore the similarities between behavioral data and model predictions, we compute the reaction to PE (participants switch or maintain their previous decision after inconsistent outcomes) as (1) maintain their decision "inconsistency/no change" (for example, after violation of trust, the participant still believe the counterpart to be trustworthy) and "inconsistency/change" (in the same example, after violation of trust, a participant retaliates) as respectively 0 and 1, and smoothed the generated vector with a window average filter with a 5 trial kernel, a computation used to probe similarities between model outputs and behavioral data (Sugrue, Corrado, and Newsome, 2004; Bayer and Glimcher, 2005; Kazuyuki Samejima, 2005).

To probe whether behavioral responses fitted our best RL model; the GL Adaptive-Belief model, we directly compared participants' behavioral choices to the outputs of the model that represented two predictions about participants' behavioral choices during the task: (1) the model predicted a strong negative prediction errors after violation of trust of a cooperative participant which should make the participant more likely to switch decision on the subsequent trial, and (2) the model predicted that the greater the value of a decision, the more likely the participant was to choose that decision.

We decided to test both hypotheses. First of all, we generated all prediction errors time-series using the different RL models on each trial for each participant and looked at these theoretical prediction errors to the smoothed ratio of "Inconsistent/No change" (i.e., when a counterpart violated participant's trust but still participant decided to trust again on the following trial) versus "and "Inconsistency/Change" (i.e., with the same example, on the following trial, the participant would have decided to switch decision). As illustrated in our picture, the predictions of the RL algorithm would correlate with the participant's actions. Particularly, as predicted by the model data, greater negative prediction errors generated by the RL model were significantly associated with the larger likelihood of participants deciding to retaliate on the next trial (average $r^2 = 0.32$; $p = 0.006$, **See figure 3.5.5**).

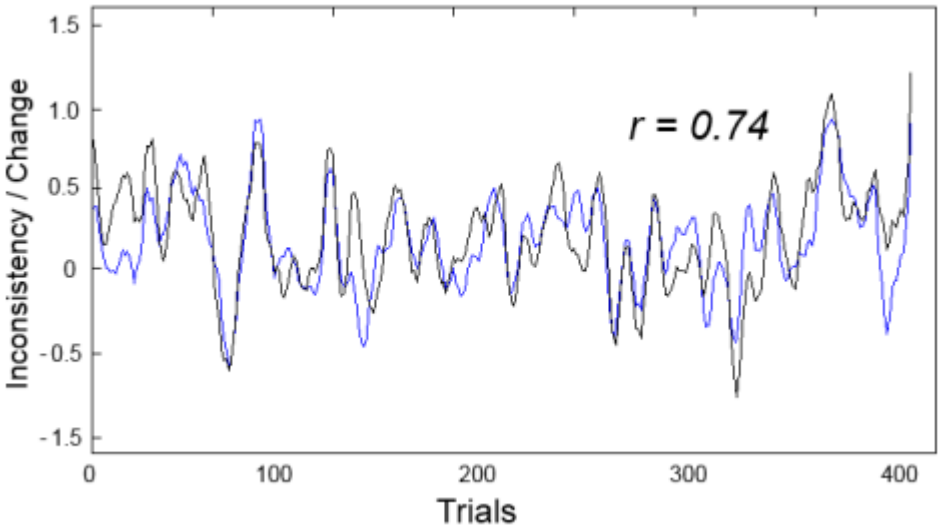


Figure 3.5.5: Outputs of the RL model (black lines) predicted subjects' trial-to-trial behavioral changes (blue lines). Results are displayed for one participant for whom the model closely fits the behavioral results. The calculated prediction error of the model on each trial closely matched the local fraction (calculated by smoothing behavioral choices, coded as 0 or 1, with a 10 trial kernel running-average filter) of participants "Inconsistent/Switch" versus "Inconsistent/No Change" on each of the relevant trials.

Indeed, further analysis revealed that participants appeared to adapt their behavior more radically when their counterpart were untrustworthy (their learning rates were higher across models, mean $\alpha_L = 0.45$) as compared to when their counterpart reciprocated (mean α_G across models = 0.031). Moreover we found little evidences supporting that reputation would only influence first impression. When only manipulating independently the way positive and negative outcomes would be updated, we found that the GL model did not fit the behavioral data more than the baseline model. However, when conceptually used in other models, free parameters for gains and losses highly improved some model-fits. In fact, the parameter $\beta_{violate}$ that was estimated in the Adaptive-Learning model (when a counterpart violates trust) was es-

sentially the highest susceptibility value for learning, indicating that participants were strongly affected by violation of trust especially from cooperative partners associated to positive reputational priors.

Finally, we explored the second prediction of the RL models: the values generated by the RL model would fit with decision to trust versus decision to not trust chosen by the participant. Our best model was used to generate values of the two options for each trial, based on each participant's peculiar history of decisions. We computed the difference between the two values at each trial and compared it with the local ratio of decision to trust versus decision not to trust (coded as 0 or 1). In complement to our previous findings, we found a significant correlation between what the model estimated and what the participants actually chose. Particularly, larger relative values of the decision not to trust versus decision to trust were associated with increased likelihood that the participants would not trust (average $r^2 = 0.41$; $p < 0.001$).

3.6 Discussion

This first study explored the behavioral and computational mechanisms underlying the decision to trust (or not) an unfamiliar person in a repeated financial task which has been explicitly proposed to operationalize the investigation of trust and trustworthiness (Berg et al., 1995). Previous studies have suggested that both prior information provided by a third party (Delgado et al., 2005; van 't Wout and Sanfey, 2008) and direct observation (King-Casas et al., 2005; Singer, Seymour, O'Doherty, Stephan, Dolan, and Frith, 2006) directly influence decisions to trust and representation of trustworthiness. For the first time in this study, we account for both variables (with and without reputation information) to probe how these variables interact in a social setting. We found that both the reputational prior information of a counterpart and subsequent experience with that counterpart synergistically impact behavioral data in this game.

3.6.1 Behavioral measures of trust

Consistent with our hypothesis, we observed that prior reputation of someone unknown directly influenced participant's initials decision to trust (or not) (van 't Wout and Sanfey, 2008). Indeed, participants were willing to take the risk of trusting these counterparts more than 70% of the time on the first rounds played with someone that has a reputation of cooperator. On the other hand, if their counterpart had a reputation for being individualistic, then participants

decided to trust them only 20% of the time. This result, at first, provides more support to the notion that social behaviors can be conveyed through reputation information and that participants are indeed sensitive to the nature of the reputation. In a similar study, Chang et al (2010) show that counterpart facial expression (positive versus negative) also influence decision to trust in RTGs. Therefore, initial knowledge provided by a third party directly influences initial judgment (Chang et al., 2010).

In this study, we found that participants were more willing to trust counterparts with a high pro-social reputation as compare to counterparts with an individualistic reputation. Therefore, we can suggest that reputational information serve as risk signals that can impact on the decision to trust and expectation of being trust in return. However, compared to pure risk, other neuroscientific studies report that unique features to the social nature of trust can be selectively manipulated, for example by introducing a hormone induction into the system (Fehr, Fischbacher, and Kosfeld, 2005). This hormone, called oxytocin, plays a role similar to a neurotransmitter in the central nervous system and influences decision to trust via the amygdala (Baumgartner, Heinrichs, Vonlanthen, Fischbacher, and Fehr, 2008).

With this first series of analysis, we also replicated results of previous neuroeconomics studies which report that people use direct observation of past events as a basis for their trustworthiness judgments (King-Casas et al., 2005).

During the course of our experiment, participants learned to trust more often counterparts that reciprocated frequently, and less often counterparts that did not. Taken together this finding suggest that decision making during repeated interactions is directly impacted by both explicit reputation priors and also by direct social signals conveyed via experience during the game. Additionally, reputation and trust seem to act synergistically to impact the observed behavior of our participants. Indeed counterparts that have a reputation for being cooperative were initially viewed as more trustworthy and then were trusted more often during the RTG in the game. This result suggests that reputational priors impact both initial judgment and experience.

3.6.2 Modeling trust

To better understand the individual learning processes underlying our behavioral findings, we modeled six possible learning processes. While there are of course many other models that could have been tested, we chose to focus on the ones that have a strong conceptual grounding. Our first analysis focused the *model-free* TD learning model, classically used in learning situation. Then, we fine-tuned this baseline model by using *model-based* RL algorithms. The

"Gain and Losses" model captures a well-known human bias: a differential sensibility to gains than losses (the prospect theory). In a third model, we tested a pure Prior-Expectation assumption that just operates on the initial value of decision to trust. Following, we tested the Prior-Decay algorithm hypothesizing that reputational prior will impact behavior responses early in the relationship, but will ultimately be overridden by experience, an assumption similar to a dual process mechanism (Poldrack, Clark, Paré-Blagoev, Shohamy, Creso Moyano, Myers, and Gluck, 2001; Frank, O'Reilly, and Curran, 2006). Then, we used a model implying that the trustworthiness representation impact on how outcomes of the interaction can be interpreted. The "Outcome-bonus" hypothesis allows that outcomes consistent with the starting trustworthiness belief will be rewarded a learning "bonus", whereas outcomes that are inconsistent (i.e. a reputed cooperative counterpart that defects) will be disregarded (Doll et al., 2009; Biele et al., 2009; Chang et al., 2010; Biele et al., 2011). Finally, we tested the Adaptive-Belief learning algorithm assuming that trustworthiness judgment is reliably used as a learning advantage in the update calculation but will eventually be updated on the basis of their counterparts' behavior after each game.

We employed a cross-validation technique to assess the best-fit of our models. We found that a simple initialization account does not fit the data as fine as the other RL models that allowed the starting judgment to impact the update function. In line with this result, we also found a behavioral interaction between reputations and trust that was not entirely driven by the first interactions but rather that lasted until the end of the relationship (**See figure 3.5.1**). Additionally, we found that our new model and the outcomes bonus model essentially provide additional confirmation that the general reputational process is not a phenomenon that is overridden by experience. Further, we found that our model that allows trustworthiness judgment to both capture reputation signal and that allow the beliefs to be updated based on the outcome of each interaction explains the behavioral data the most. Notably, our proposed model predicted two effects that are confirmed by our behavioral data.

Our first finding suggests that the Adaptive-Belief algorithm highly predicted the behavior of our participants and most specifically, the fact that toward the end of a RTG, participants decided to trust the most counterparts that reciprocate with a high probability and the least to counterparts that did not reciprocate frequently. Our results confirmed this hypothesis (figure reference) and in particular, we report the exact behavior for the data in all the manipulated experimental conditions and the data predicted by the RL model.

We found that the Adaptive-Belief model predicts that trustworthiness judgment would be first influenced by the nature of the reputation information and will change over time based on actual experiences. Because the model-predicted decision to trust switch more rapidly when inconsistent behavior was observed from counterparts with reputations, the predictions were

also better able to predict behaviors than models that do not account for interaction between trust and reputation. These results support and extend previous findings, which have found that, although instructions alert decisions in uncertain environment, experience overwhelms explanation in risky choice (Barron and Erev, 2003; Hertwig, Barron, Weber, and Erev, 2004; Jessup, Bishara, and Busemeyer, 2008). While this notion of prior expectation being dynamically updated with experience is certainly not new, this study proposes for the first time a computational model of this effect in a social context and provides support for its validity.

3.7 Conclusion

Our first study integrates methods and theories from different disciplines: psychology, neuroscience and economics with the framework of reinforcement learning. Our goal is to gain a greater understanding of how high-level social cues such as reputation information and trustworthiness are interpreted and used in an iterative social exchange. Our results suggest that decision to trust with or without reputational prior engage different cognitive processes. While decision to trust without reputational priors is based on actual experiences through repeated interactions, decision to trust when reputation prior are available is determined by the nature of the reputation and then dynamically updated based on direct observation.

In this first study, we propose a new computational RL model with an interdisciplinary approach to conceptualize the notion of learning to trust with or without reputation information. This model proposes a novel approach to bridge the division between theoretical models and empirical evidences of trust and reputation in the social decision-making literature (Jessup et al., 2008). Overall , our study provides new evidence supporting a growing literature involved in the neural computations underlying adaptive social learning (King-Casas et al., 2005; Behrens et al., 2008; Biele et al., 2009, 2011; O'Doherty et al., 2007) and decision to trust mechanisms (Krueger et al., 2007; van 't Wout and Sanfey, 2008). Finally, this first study illustrates the importance of prior-expectation in social decision-making that can be used as a risk belief (i.e. the likelihood for reciprocity).

Your theory is crazy, but it's not crazy enough to be true.

Niels Bohr

4

Neural Correlates of Trust and Reputation[1]

4.1 Background

Trusting others involves risk and uncertainty: people invest a form of good (i.e. money, work, time etc.) in interactions that can yield a profit or a loss, depending on whether others hold to their end of the bargain (Coleman, 1994). Critically, when others are not contractually committed to doing so, they may be untrustworthy for their own benefit and harm the person that initially placed trust in them (Berg et al., 1995). In financial transactions, investors should then either anticipate this, and not invest money to begin with, or develop efficient strategies to estimate the trustworthiness of others (Camerer and Weigelt, 1988). Experiments with repeated Trust Games (RTGs) allow to empirically observe trust-based dynamics (Chang et al., 2010). Neuroimaging studies employing RTGs have shown that, when no prior information on transaction partners is available, the brain's reward circuitry is involved in learning about their type (i.e. their level of trustworthiness), based on the outcomes of previous trust-based interactions (King-Casas et al., 2005). Indeed, reward-related brain re-

[1]Parts of this Chapter have been taken from Fouragnan et al. (2013). *Reputational Priors Magnify Striatal Responses to Violations of Trust*, The Journal of Neuroscience

gions have been found to respond positively to trustworthiness and negatively to violations of trust (Krueger et al., 2007; Phan et al., 2010; Long, Jiang, and Zhou, 2012). We refer to this as *"interaction-based"* learning. However, a second important alternative for investors to efficiently engage in financial decisions is to rely on priors provided by a third-party. Such priors may affect the way agents evaluate the outcomes of transactions and thus how they learn about the type of their counterparts. We refer to this as *"prior-based"* learning. For example, in web-based transactions, which are increasingly used, investors interact with complete strangers and rely on available reputation priors (e.g., reports on previous transactions, customer reviews etc.) to predict expected returns and potential risks associated with investments (Kim, 2009). However, while the neural correlates of interaction-based learning to trust have been largely explored, only few studies have investigated the neural bases of trust when reputation priors are provided (Delgado et al., 2005; Stanley et al., 2012). No studies on date have directly compared the two forms of trust-based decision making within the same experiment. To confront this issue, we conducted a functional magnetic resonance imaging (fMRI) experiment in the attempt to characterize the neural activation patterns related to trust-based decisions during RTGs. Two situations were analyzed and compared, one in which we provided information about the social attitude of counterparts (i.e. reputational priors), and one in which no such information was provided. Furthermore, in contrast to a previous neuroimaging study on the same issue (Delgado et al., 2005), we also manipulated the actual level of trustworthiness demonstrated by counterparts during an RTG, such as to make it consistent with the provided priors. Finally, we used standard fMRI analysis, *model-free* and *model-based* reinforcement learning (RL) models to approach the problem of social learning and reputation effects. Our main goal was to assess whether and how reputation priors affect RL mechanisms at both the behavioral and neural level.

4.2 Experimental Design and Methods

4.2.1 Participants

Twenty male participants (mean age, 29.5 $\pm$ 3.53 years) took part in the fMRI experiment; two were removed from the analysis for excessive head movement (See fMRI analysis). All of them were healthy, gave written informed consent, had normal or corrected-to-normal vision without any history of psychiatric, neurological, or major medical problems, and free of psychoactive medications at the time of the study. Participants were told that the experiment aimed at studying decision making in a social context, that they would receive a compensation of 15 Euros/hour and that the money gained in ten randomly extracted trials would be added

to their compensation. The study was approved by the local institutional ethical committee of the University of Trento.

4.2.2 Task

The experimental task was based on the Trust Game (TG) (Berg et al., 1995). In one round of our task, each participant played as "investor" with an anonymous counterpart as "trustee". Both players were endowed with 1 euro before starting a round composed of 2 stages (**See figure 4.2.1**): in stage 1 the participant decided whether or not to share his euro with the trustee. If he decided to share, the euro was multiplied by 3 by the experimenter before being allotted to the trustee. In stage 2 the response of the trustee could be to either equally share his money with the investor (1/2 of 4 euros = 2 euros) or keep his money and return nothing. It follows that if the investor invested and the trustee reciprocated, both players were better off than if the interaction has not occurred at all. However, investing was risky, as if a trustee returned nothing, the investor incurred a loss.

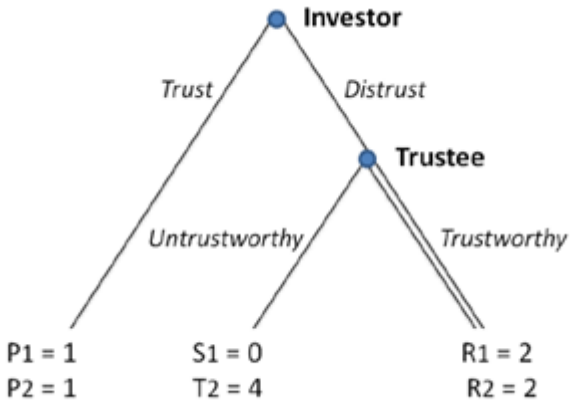


Figure 4.2.1: Experimental design. One round of the two-player repeated trust game (RTG). P1 is the payoff of the participant, who always plays as investor; P2 is the payoff of his counterpart, who plays as trustee. Before each round both players are endowed with 1 euro. The participant moves first and chooses either to "keep" (Trust) or "share" (Distrust) If he keeps, both players maintain their initial endowments. If he shares the participant's endowment is multiplied by 3 and passed to the counterpart. The trustee then decides whether to share in turn by returning 2 euros (Trustworthy), or to keep by returning nothing (Untrustworthy). RTGs consisted of several consecutive rounds with a same counterpart. Participants played with many different counterparts and were told that their counterparts had already made their choices.

We used a repeated version of this TG (RTG), which consisted in a series of consecutive TG rounds with a same counterpart. However, this alters the nature of the single-shot TG, as RTGs allow for additional strategic maneuver. For instance, investors tend to invest more (and

trustees to reciprocate) in initial rounds of RTGs, than in final rounds or single shot games (Isaac, McCue, and Plott, 1985). For similar reasons, both parties may strategically punish (by not investing) if they believe this might incentivize uncooperative counterparts to review their strategies in future rounds. Our study intended to minimize the strategic component of trust-related behavior; hence our version of the game differentiated from the typical repeated TG in few but important respects.

1. Subjects were informed that trustees had already made their choices, which thus wouldn't have been affected by those of the participant. In other words, participants knew that counterparts were not interactive. This feature should have eradicated any strategic component usually present in RTGs. In reality, the trustees were computer simulations and they reciprocated an investment with fixed probabilities unknown to participants.

2. Another feature was also adopted to make learning independent on participants' actions. In traditional RTGs, when an investor does not trust, the round ends and nothing is learned about the behavior of counterparts. In our study, on the other hand, participants learned about the trustees' choices even when they invested nothing. This adjustment enabled to keep the amount of feedback fixed (regardless the choice of participants), thus allowing us to compare learning mechanisms between conditions.

3. Finally, to further reduce strategic reasoning, participants did not know how many games composed each RTG with a given trustee but only that RTGs were consecutive and if they were not paired with the same trustee twice in a row, then they would have never encountered the counterpart again. Specifically, we fixed a constant probability of $1/3$ to continue the game with a same counterpart; this resulted in a minimum of one and a maximum of eight games with a same trustee.

Then, each trustee was introduced with a picture of his face before a RTG began (**See figure 4.2.2**). The association between pictures and RTGs was randomized, as was the order of RTGs. To reduce facial information extraction and gender attraction, we assembled a database of colored pictures from 20 to 60 years old Caucasian men (mean age and SE: 34.05 ± 11.19) controlled for attractiveness, emotion and racial traits. 128 pictures were selected and used with authorization from the FERET database of facial images collected under the FERET program (Phillips et al., 2000). The words "trust" or "trustworthy" were never mentioned.

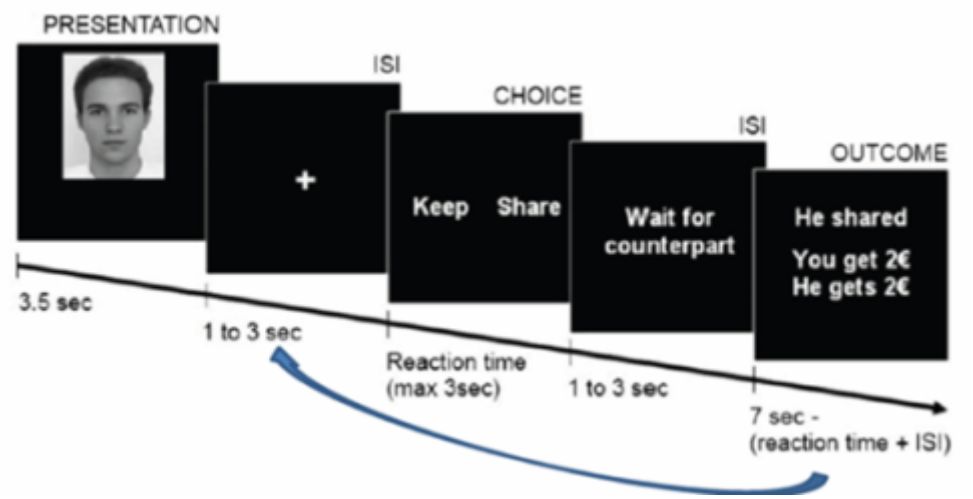


Figure 4.2.2: Timeline of the first RTG round. Presentation: Face of the counterpart (with a prior or no-prior) was displayed for 3.5 s, and only presented for the first round of an RTG. Fixation: Fixation cross was presented during a jittered inter-stimulus interval (ISI). Choice: Participants made their choice by pressing "Keep" or "Share". Delay: ISI corresponding to the (simulated) decision of the counterpart. Outcome: Outcome of the game and the payoffs of each player.

4.2.3 Experimental Setup

Cooperative trustees would reciprocate 80% of the times, while individualistic counterparts would defect 80% of the times (though participants were not informed of such contingencies). The distinction between types furthermore allowed confronting the cases in which trustees behaved consistently ("Cons") or inconsistently ("Incons") with their types.

The second key feature of our study was whether or not a reputation prior was provided (**See figure 4.2.3**). In the prior-condition, half of the cooperative and half of the individualistic trustees were flagged, respectively by a circle and a triangle. These cues signalled their "reputation". Specifically, participants took part in the Social Valuation Orientation (SVO) (Messick and McClintock, 1968; Van Lange, 1999) and were told that the distinct cues were based on the trustees' scores for the same task. This task distinguishes between different types of social value orientations (e.g., cooperative or individualistic). The main difference between each category is the extent to which one cares about own payoffs and that of the others in social dilemma situations.

Finally, for the remaining half of the counterparts, no prior information was provided (no-prior condition).

To ensure no difference in learning scheme in each of the four conditions (Prior Cooperative, Prior Individualistic, No-Prior Cooperative, No-Prior Individualistic), RTG length and share/keep schedules within each RTG were counterbalanced.

4.2.4 Procedure

Participants received written instructions, took part in a simplified version of the SVO task, and completed a 20 min RTG practice session (20 trials). The experiment was implemented using Presentation software (version 0.70).

A first key manipulation was that trustees were divided into 2 predefined types: they could be either "cooperative" or "individualistic" (**See figure 4.2.3**).

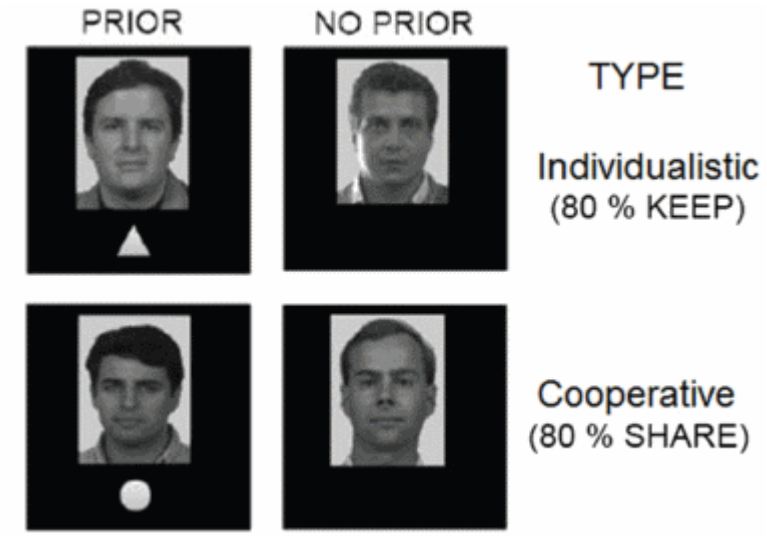


Figure 4.2.3: Experimental conditions. Two conditions were adopted: 1) the "type" of counterpart, and 2) the presence vs. absence of "reputational priors". Types: counterparts could be either "cooperative" or "individualistic" in their (simulated) behavior in RTGs; the former shared and the latter kept in 80% of RTG rounds. Reputational priors: participants were told that cues indicated whether the current counterpart had obtained a high or low score in a social orientation task (triangles indicated low scores, circles indicated high scores). Such priors reliably differentiated between the 2 counterpart types.

In the scanner, subjects completed 356 trials (89 for each condition: Prior Cooperative, Prior Individualistic, No-Prior Cooperative, No-Prior Individualistic), divided in four runs of 20 min. **Figure 4.2.2** shows the time line of the first trial of an RTG. Each RTG started with a 3.5 s display of the face of the trustee (which, only in "prior" conditions, was flagged with a reputational cue). This was followed by a fixation cross and then by a "decision screen," which required participants to choose between one of two options, labeled "share" or "keep." After making their choice, participants waited a jittered interval before an "outcome screen" appeared, displaying the trustee's choice and the corresponding payoffs to both players. For those trials in which participants chose to keep, the outcome screen was still shown.

4.2.5 Analysis

Behavioral data analysis

Behavioral data were analyzed using Stata Statistical Software version 9.2 and the R environment (Development Core Team, 2008). A two-way repeated measure ANOVA was performed to identify differences between conditions for each variable of interest (e.g., decision to trust, payoffs made in each condition). Next, we computed regression analyses using mixed-effects linear models (MEL), in which participants were treated as random effects and hence were allowed to have individually varying intercepts. Parameter estimates (b), *SE*, t values and p values were reported.

RL models

Model 1: *model-free* temporal-difference learning.

We first used a "model-free" temporal-difference (TD) (model 1) learning algorithm (Rummery and Niranjan, 1994; Sutton and Barto, 1998), which assumes that agents are initially unaffected by the presence of priors, but that, as trials with a counterpart unravel, they may update reward values differently when priors are available as opposed to when they were not available. Participants would sample the reward probability of two choices (Keep or Share) in the Cooperative and Individualistic conditions. We then hypothesized that participants would obtain reliable expectation of these conditions updating the estimated value of each choice with a discounted "step-size." Thus the stochastic prediction error δ, based on the Rescorla–Wagner learning rule (Rescorla and Wagner, 1972) was computed as follows:

$$\delta_t = R_t - V_{C,t} \tag{4.1}$$

where R is the payoff obtained at time t, when choosing an option C at time t or $t+1$, and V is the value of each choice Share or Keep in each trial. In addition to this, the following learning rule differentially updated the stochastic prediction error in the Prior (P) and No-Prior (NP) conditions:

$$V_t + 1 = V_t + a^P.\delta^P_{(C,t)} + a^N P.\delta^{NP}_{(C,t)} \tag{4.2}$$

The degrees in which δ^P and $\delta^N P$ influence the new action value are weighted by two learning rates, αP and αNP, where $0 < a^P, a^N P < 1$.

Model 2: model with separate expectations for positive or negative priors.

Additionally, given the results found in the previous study (**See Chapter 3**), we hypothesized that, in the Prior condition, participants have "optimistic" or "pessimistic" expectations, at the beginning of the game due to the presence of a positive (P^+) or negative Prior (P^-), respectively (Biele et al., 2009, 2011; Wittmann et al., 2008). Thus, the values of initial choices when playing with a Cooperative or Individualistic counterpart in the prior condition were formally defined as:

$$V^{Prior+}_{(C,0)} = g^{P+}.\mu\theta_{Prior+}.N \tag{4.3}$$

$$V^{Prior-}_{(C,0)} = g^{P-}.\mu\theta_{Prior-}.N \tag{4.4}$$

where g^{P+} and g^{P-} are equal to 1 when playing with a counterpart with a positive or negative prior, respectively; and 0 otherwise. θ_{P+} and θ_{P-} are free parameters capturing the optimistic or pessimistic impact of the priors expectation, μ is the expected payoff from choosing randomly among all options, which serves as a normalization constant and N is the number of trials experienced in the learning condition, which is a scaling factor, allowing for the comparison between an expected decision value and the outcome of the decision. On the other hand, in the no prior condition, only one parameter weighted the initial expected value of choices, $V^{NP}_{(C,0)}$.

The Softmax function was then used for the two models to determine the probability of choosing a given choice option given the learned values as follows:

$$P_{(share)t} = \frac{exp^{(\frac{V(share)_t}{\tau})}}{exp^{(\frac{V(share)_t}{\tau})} + exp^{(\frac{V(keep)_t}{\tau})}} \tag{4.5}$$

where τ is called a temperature parameter. For high values of τ, all actions have almost the same probability (i.e., choices are random), while for low τ the probability of choosing the action with the highest expected reward ($V1 > V2$) is close to 1.

To generate model-based regressors for the imaging analysis, both learning models were simulated using each subject's actual sequence of rewards and choices to produce per-trial, per-subject estimates of the initial values V_t and error signals δ_t (Morris et al., 1996; Wittmann et al., 2008). All parameters of interest were implemented in MATLAB R2009 and were estimated using the negative log likelihood of trial-by-trial choice prediction. Model comparisons were performed with the Bayesian Information Criterion, the pseudo r^2 value using the Log likelihood of a random distribution, and tested with the likelihood ratio test.

4.2.6 fMRI Data Acquisition and Analysis

fMRI data acquisition

A 4T Bruker MedSpec Biospin MR scanner (CiMEC, Trento - Italy) and an 8-channel birdcage head coil were used to acquire both high-resolution T1-weighted anatomical MRI using a 3D MPRAGE with a resolution of 1 mm3 voxel and T2*-weighted Echo planar imaging (EPI). The parameters of the acquisition were the following: 34 slices, acquired in ascending interleaved order, the in-plane resolution was 3 mm3 voxels, the repetition time 2 sec and the echo time was 33ms. For the main experiment, each participant completed 4 runs of 608 volumes each. An additional scan was performed in between two different runs in order to determine the point-spread function that was then used to correct the known distortion in a high-field MR system.

Preprocessing

The first five volumes were discarded from the analyses to allow for stabilization of the MR signal. The data were analyzed with Statistical Parametric Mapping 8 software (SPM8; Welcome Department of Cognitive Neurology, London, UK) implemented in MATLAB R2009 (MathWorks). We used SPM8 for the preprocessing steps. Head motions were corrected using the realignment program of SPM8. Following realignment, the volumes were normalized to the Montreal Neurological Institute (MNI) space using a transformation matrix obtained from the normalization process of the first EPI image of each individual subject to the EPI template. The normalized fMRI data were spatially smoothed with a Gaussian kernel of 8 mm (full-width at half-maximum) in the (x, y, z) axes. Imaging data for participants with head motions exceeding one voxel (3 mm) in transition and 3° in rotation were discarded (Eddy et al., 1996). We also used the xjView package and MRICron to create the pictures presented in the results (version 1.39, Build 4).

fMRI analysis

GLM 1a and b.

Our first analysis considered the main effect of the presence or absence of reputation priors when a new counterpart is presented for the first time. We used a general linear model (GLM), estimated in three steps: (1) first, an individual blood oxygenation level-dependent

(BOLD) signal was modeled by a series of events convolved with a canonical hemodynamic response function. The regressors representing the events of interest were modeled as a box-car function with onsets at the beginning of each RTG ("Pre") and durations of 3.5 s. For GLM 1a, regressors represented trials in which priors were provided ("Prior_Pre") and no priors were provided ("NoPrior_Pre"). For GLM 1b, regressors represented trials in which priors were provided for a cooperative counterpart ("Prior+_Pre"), priors were provided for individualistic counterparts (Prior-_Pre), and no priors were provided (NoPrior_Pre). For t contrasts, we then computed first-level one-sample t tests comparing trials with and without priors on the basis of the GLM 1a. (2) We then analyzed second-level group contrasts. Our fMRI results were initially thresholded at $p < 0.001$ uncorrected and were subsequently cluster-thresholded at $p < 0.05$, familywise error (FWE). All reported coordinates (x, y, z) are in MNI space. Anatomical localizations were performed by overlaying the resulting maps on a normalized structural image averaged across subjects, and with reference to an anatomical atlas. (3) Finally, we used the MarsBaR toolbox from SPM8 to perform functionally defined (based on the averaged parameter estimates in the cluster found with GLM 1b) region of interest analysis (ROI) and compute percentage signal changes.

GLM 2: *Prior-model* fMRI analysis.

A second GLM model uses the estimates of the best-fit RL algorithm described in **equations 4.3** and **4.4**.

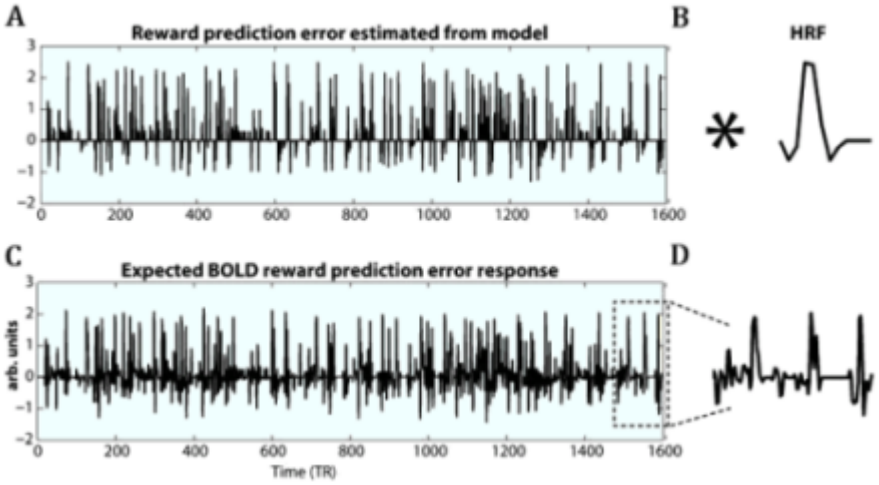


Figure 4.2.4: Illustration of a whole-brain analysis of Trial-to-Trial estimates of Prediction Error δ_t from the Prior-Expectation model from simulated data. Example of one scan. Convolution of parametric changes in estimated PE for each decision on each outcome (top plot - left) convolved with a canonical hemodynamic response function (top plot - right) produced individual participant relative prediction-error regressors (bottom plot - right).

This model still focused on the distinction between prior and no-prior conditions but additionally separated between two phases of the RTG: the decision phase and the outcome phase. This allowed us to assess how the impact on the BOLD signal of priors was parametrically modulated by two behaviorally estimated learning measures (from model 2, **See figure 4.2.4**): (1) at time of choice, the parameter Q_t, weighted the value of options, on a trial-to-trial basis, depending on RTG history; (2) while δ_t scaled outcomes on the basis of their estimated prediction error. Parametric modulation were applied to the magnitude of the stick functions. We performed this analysis at the individual level and ran group statistics, taking individual participants as random effects. We then focused on a subset of our resulting brain regions on the basis of effect strength ($p < 0.05$, FWE corrected). Specifically, averaged parameter estimates were extracted from bilateral caudate (MNI coordinates: $(-14, 20, 2)$ and $(12, 18, 6)$), separating between prior versus no-prior contexts.

GLM 3: violation of trust.

In a third GLM we differentiated between consistent (Cons) and inconsistent (Incons) outcomes. We classified consistent outcomes as those rounds in which either (1) participants had kept with individualistic counterparts that defected (Cons−) (distribution of trials: M = 57 ± 3) or (2) they had shared with a cooperative counterpart that reciprocated (Cons+) ($M = 56 \pm 4 trials$); inconsistent outcomes, on the other hand, occurred when either (3) participants had kept with an individualistic counterpart that reciprocated (Incons−) (M = $14 \pm$ 4 trials) or (4) they shared with a cooperative counterpart that defected (Incons+) (M = 15 $\pm$ 4 trials), and who thus "violated" their trust.

GLM 4: Cooperative vs. Individualistic.

In a fourth analysis, we were interested in the brain regions that were differentially solicited when participants were playing against individualistic trustees as compare to playing with cooperative trustees. We constructed a a new GLM for each participant in which the effects of playing with an individualistic counterpart ((Individualistic_Prior) + (Individualistic_No Prior)), playing with a cooperative counterpart ((Cooperative_Prior) + (Cooperative_No prior)) and parameters of head movements served as regressors.

Functional connectivity analysis (PPI).

To explore the interplay between the caudate and other brain regions following violations of trust (Incons+), we assessed functional connectivity using psychophysiological analysis (Friston, Buechel, Fink, Morris, Rolls, and Dolan, 1997; Cohen et al., 2005), which compares the pattern of activity of a seed region to every other regions of the brain. We took the bilateral caudate resulting from the reported GLM 3 (Cons > Incons) as seed regions, as these areas showed highest sensitivity to violations of trust ($t = 6.78$, $p < 0.05$, FWE). Then, we created three regressors: (1) the caudate time course (physiological regressor), (2) an event-related regressor that distinguished between violations of trust in the prior and no-prior conditions (with a boxcar function ranging from the beginning of the outcome phase until the end of the interstimulus interval; ISI), and (3) the interaction term. Additionally, we also conducted a correlation analysis between the retaliation rate for each subject (measured by the percentage of choices to keep after violation of trust when playing with a cooperative partner) and the parameter estimates in left ventrolateral prefrontal cortex (vLPFC) (MNI −40, 42, 4) across subjects. Finally, to examine how striatal responses to violations of trust were related to learning, we plotted individual parameter estimates against the individual learning rates (estimated with model 2 described above).

4.3 Behavioral Measures and Computational Results

4.3.1 Behavioral data

Our main goal was to determine whether reputation priors influence initial expectations and decisions in the games, and subsequent learning mechanisms. A repeated measure two-way ANOVA was performed using type of counterpart (cooperative or individualistic) and prior condition (prior or no-prior) as within participant factors.

The percentage of decisions to share was significantly higher with cooperative counterparts ($M = 71.77$, $SE \pm 4.03$) than with individualistic counterparts ($M = 27.34$, $SE \pm 3.71$; $F_{1, 17} = 174.01$, $p < 0.001$). The results also showed a significant interaction effect of prior with type of counterpart ($F_{2, 35} = 30.87$, $p < 0.001$). Post hoc tests (t-tests Bonferroni corrected) indicated that participants decided to share with cooperative partners more when provided with a prior ($M = 81.09$, $SE \pm 4.78$) than when priors weren't provided ($M = 62.45$, $SE \pm 5.81$; $t = 5.89$, $p < 0.001$), whereas they decided to share with individualistic counterparts less in the prior ($M = 18.37$, $SE \pm 4.66$) than in the no prior condition ($M = 36.3$, $SE \pm 5.05$; $t = 4.23$, p

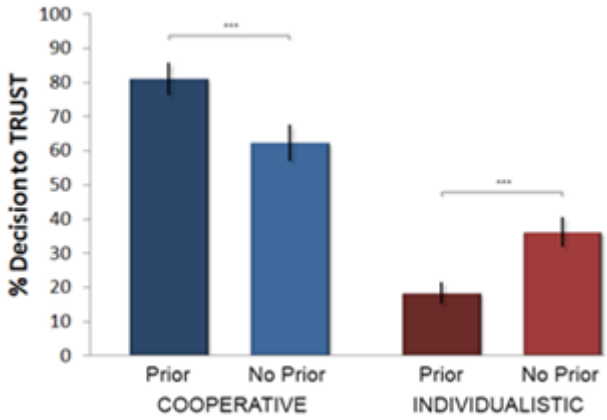


Figure 4.3.1: Behavioral results. Average percentage of decision to trust across conditions. Mean ± standard error of participants' decision to trust (share) are broken down for trustee's type (Cooperative/Individualistic) and prior condition (Prior/No Prior); *** p<0.001. Priors enabled participants to match (on average) their choices with the counterpart's level of trustworthiness.

< 0.002, **See figure 4.3.1**).

When payoffs are analyzed with type of counterparts and prior condition as within-subject variables, we found that payoffs were significantly higher when playing with cooperative counterparts (M = 1.43, *SE* ± 0.13) than individualistic counterparts (M = 0.94, *SE* ± 0.11; F1, 17 = 138.32, p < 0.001) and significantly higher in the prior condition (M = 1.20, *SE* ± 0.10) than the no prior condition (M = 1.08, *SE* ± 0.06; F1, 17 = 28.98, p < 0.001, **See figure 4.3.2**).

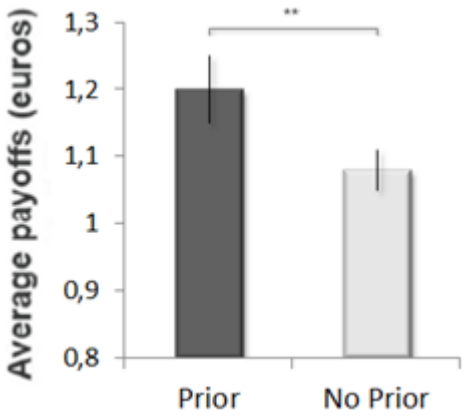


Figure 4.3.2: Average payoffs in the Prior and No-Prior conditions. Average payoffs ± standard error (in euros) in Prior/No Prior conditions. When priors are available, participants significantly earn more when they adjust their choices to counterparts' types; ** p<0.01.

In order to examine the effect of the prior condition, the trustees' type, the order of the repeated game and the interactions of such factors on the decision to share (binary dependent variable), we performed regression analyses using mixed-effects linear (MEL) models. The re-

sults revealed that participants shared with cooperative counterparts more often as compared to individualistic counterparts (b = 1.29 ($SE \pm 0.08$), t = 15.8, p < 0.001); shared less when they did not receive priors (b = - 1.09 ($SE \pm 0.09$), t = - 12.1, p < 0.001); and shared less over time (b = - 0.12 ($SE \pm 0.02$), t = - 6.81, p < 0.001). These results suggest that participants took into account reputation priors and played according to the counterpart's level of trustworthiness. Instead, when priors were not available, participants learned counterparts' types on the basis of their actions. Interestingly, we found an interaction effect between the trustees' type and the prior condition (b = 2.27 ($SE \pm 0.13$), t = 17.39, p < 0.001). These results indicate that the difference between prior and no prior conditions was greater when playing with a cooperative than with an individualistic counterpart. Furthermore, even though participants in the no prior condition adjusted their decisions to their counterparts' type over rounds, they still shared with cooperative counterparts less than when they had priors (**See figure 4.3.3**).

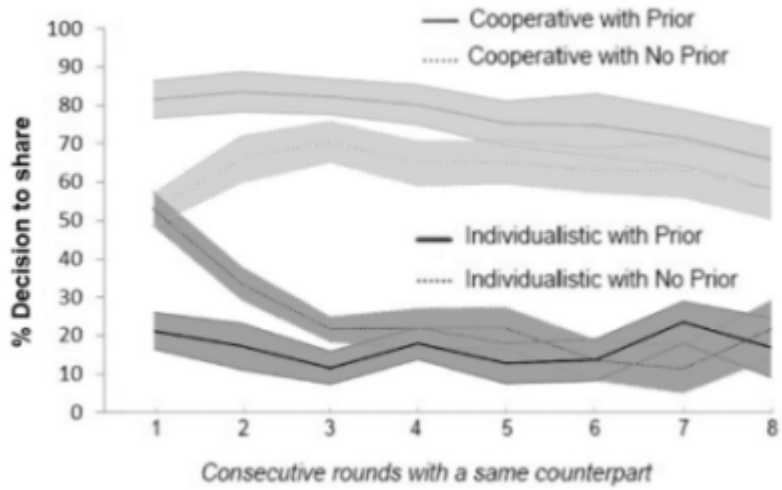


Figure 4.3.3: Learning dynamics across RTG rounds. Average percentage of the decision to trust for each round when playing with a "cooperative" vs. "individualistic" counterpart, and when priors were present vs. absent. When participants know nothing of their counterparts they tend to randomize between trusting and not trusting during initial rounds and adjust their choices to their counterparts' type in succeeding rounds. On the other hand, when priors are present, participants tend to rely on them already from early rounds. Standards errors are the shaded areas bellow and above the main curves.

Post hoc t-test revealed that, in the no prior condition, in rounds when cooperative counterparts kept, participants subsequently kept more (Mean percentage of decisions to keep = 0.48,$SE \pm 0.019$), whereas they persisted in sharing in the prior condition (M = 0.2, SE $\pm$ 0.015; t_{17} = - 4.99, p < 0.001), (**See figure 4.3.4 A**). Similarly, when individualistic counterparts shared in a round, participants subsequently shared more when not provided with a prior (Mean percentage of decisions to share = 0.34, $SE \pm 0.015$) than when given a prior (M = 0.21,

$SE \pm 0.009$; t_{17} = - 4.783, p < 0.001, **See figure 4.3.4 B**).

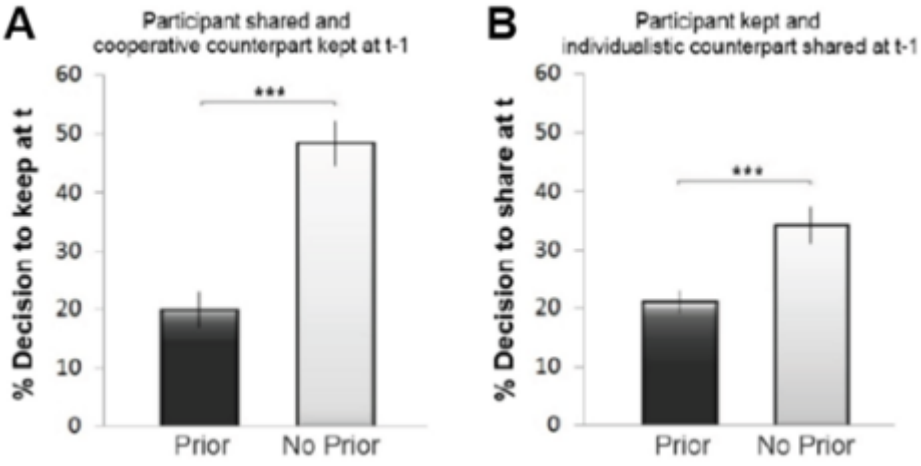


Figure 4.3.4: A. Choices following unexpected behavior of cooperative and individualistic counterparts. Average ($\pm$ standard error) of percentage of "keep" choices in prior vs. no prior condition at time t, following rounds in which participants shared and a cooperative counterpart violated their trust by deciding to keep (at t-1). Decisions to Keep at time t (i.e., retaliation) was less frequent when priors were available. B. Percentage of "share" choices (at t) following rounds in which participants had kept and an individualistic counterpart has shared (at t-1).

4.3.2 Results from learning models

A likelihood ratio test revealed that the Prior model (model 2) with separated expectations for cooperative and individualistic counterparts (Prior mode) performed better than the classical TD learning model (model 1) (p < 0.001) (Additional statistics are reported in **Table 4.3.1**).

The best-fitting parameters are shown in **Table 4.3.2**. For these parameters, we found that the average learning rate estimated from trials in the No Prior condition, a_{NP}, was significantly higher than the average learning rate estimated from trials in the Prior condition δ_t ($t_{17} = 2.29; p < 0.05$).

We also found that the initial value in the Cooperative Prior condition, $V^{P+}(0)$ was significantly higher than the initial value in the No Prior condition $V^{NP}(0)$ ($t_{17} = -2.82; p < 0.001$), and the initial value in the Individualistic Prior condition, $V^{P-}(0)$ ($t_{17} = -3.07; p < 0.001$). There was no significant difference between the initial value in the Individualistic Prior condition, $V^{P-}(0)$ and the initial value in the No Prior condition $V^{NP}(0)$,($t = 0.46$). Finally, we found that the average learning rates estimated for each participant when they kept was higher

Learning model comparison		
	Classical model-free TD learning model	Prior+ and Prior- expectations RL learning model
BIC	7619	6460
Log Likelihood	-3809	-3230
Pseudo r^2	0.14	0.273

Table 4.3.1: Learning model comparison: Bayesian Information Criterion value (BIC), Log likelihood and the Pseudo r squared suggest that the Prior+ and Prior-expectations TD learning model fits the observed behavior better the other TD learning models.

Parameter estimate for best behavioral model, depicted as mean ± SE		
	Mean	SE
Learning rate Prior condition δ_P	0.3373	± 0.0456
Estimates for Cooperative counterparts	0.327	± 0.0424
Estimates for Individualistic counterparts	0.3475	± 0.0398
Learning rate No Prior condition $\delta_N P$	0.5075	± 0.0689
Estimates for Cooperative counterparts	0.4686	± 0.0701
Estimates for Individualistic counterparts	0.539	± 0.0599
Estimates learning rates for Invest trials (participants shared)	0.3845	± 0.0459
Estimates learning rates for Non-Invest trials (participants kept)	0.4603	± 0.0476
Softmax inv. Temp Betha τ	4.7769	± 0.3149
Initial value Cooperative Prior condition $V_{P^+}(0)$	1.3814	± 0.1031
Initial value Individualistic Prior condition $V_{P^-}(0)$	0.9838	± 0.1055
Initial value No Prior condition $V_{NP}(0)$	1.0641	± 0.1260

Table 4.3.2: Learning best model Averaged best-fitting parameter estimates (across subjects) SE

($M = 0.46, SE \pm 0.04$) than when they shared ($M = 0.38, SE \pm 0.048; t_{17} = -2.27, p < 0.05$,**see Table 4.3.2**).

4.4 fMRI Results

4.4.1 GLM1 Effect of prior at time of counterpart presentation

In a first brain imaging analysis, we investigated whether it was possible to differentiate the condition in which participants received prior knowledge of the Trustee's reputation or not based on their brain activation pattern at the beginning of each new multi-round game. This period is of particular interest because, as we showed in the behavioral results section, our participants were able to anticipate the Trustee's strategy and adjust their own decision to this strategy before starting each new multi-round game when they had received prior reputation-related information (i.e., they chose to place trust or not in their counterpart with a probability close the counterpart probability of reciprocation). Thus, this period took place at a time point when the decision to trust (or not) could be implemented when the prior was provided. The contrast ("Prior_Pre" > "NoPrior_Pre") (see Methods, GLM1a) revealed differential activity in the mPFC (0, 62, 31), to the presence vs. absence of any priors when new counterparts were presented (t = 8.26; p < 0.05 FWE cor.)(**See figure 4.4.1** and **Table 4.4.1**).

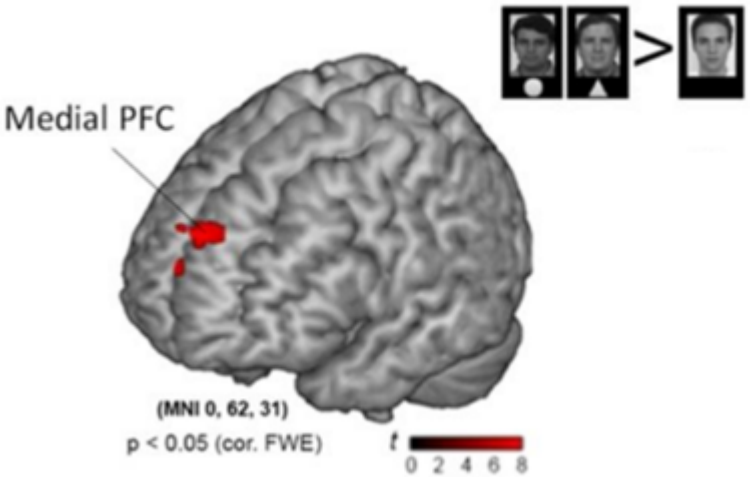


Figure 4.4.1: mPFC encodes reputational priors when a new counterpart is first presented. Random Effect Analysis. When contrasting (Prior) > (NoPrior) conditions at time of counterpart presentation, activity in the medial prefrontal cortex survived FWE correction, p < 0.05.

Further functional ROI analysis, based on GLM1b, qualified this activation pattern as responding with increased activity to the presence of priors, regardless of their nature (positive

or negative), and decreased activity to their absence (**See figure 4.4.2**). The opposite contrast ("NoPrior_Pre" > "Prior_Pre") revealed activity in bilateral anterior insula (-36, -4, 15), t = 3.91; p < 0.001 unc., and (38, 3, 10), t = 3.45; p < 0.002 unc.).

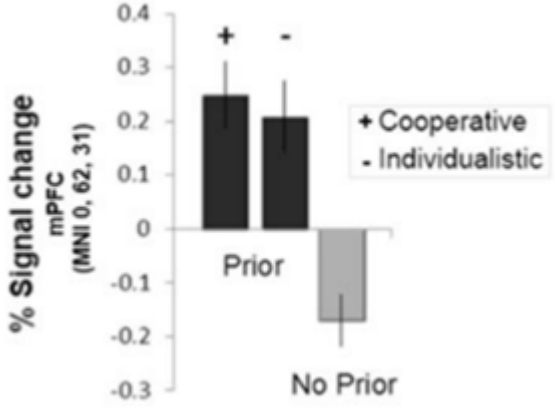


Figure 4.4.2: Functional ROI Analysis in mPFC. Functional ROI analyses further revealed percentage signal changes in the medial prefrontal cortex MNI (0, 62, 31). The figure shows an increased activity when priors were present, regardless of their type, and decreased activity when there were no priors.

4.4.2 GLM2 Effect of prior at RTG choice

Applying parametric analysis (**See Materials and Methods**, GLM 2 *model-based* fMRI analysis) to the functional MRI data, we focused on trial-to-trial weights on decision values as represented by per-trial Q_t estimate amplitude. We found that decision value estimates were correlated with neural activity in a network consisting of the mPFC (−2, 64, 10) and the dorsolateral prefrontal cortex (dLPFC) (−38, 38, 32), surviving p < 0.05, FWE corrected (**See figure 4.4.3** and **Table 4.4.1**). These two regions reflected the contributions of prior's valence (positive or negative) to the pattern of activity related to the decision to trust (**See figure 4.4.4**). Moreover, the difference at a neural level between prior and no-prior condition was greater when playing with a cooperative counterpart compared with an individualistic counterpart. This is consistent with the observed behavioral asymmetry of the effect of priors between cooperative and individualistic conditions.

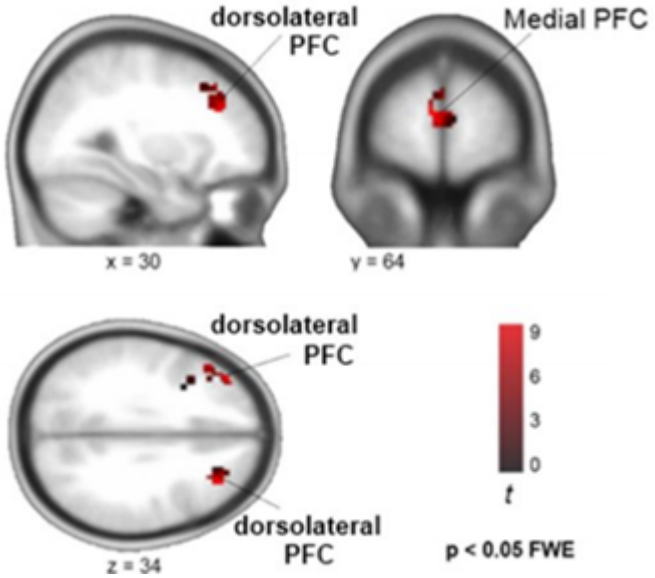


Figure 4.4.3: Brain regions parametrically correlated with the estimated "optimistic" and "pessimistic" decision value from the Prior model. Random effect fMRI analysis. To look for neural correlates of value signals (Q_t) at time of choice, we entered the trial-by-trial estimates of the values of the two stimuli ("Share" and "Keep") into a regression analysis against the fMRI data. We found enhanced activation in mPFC and dLPFC, surviving FWE correction, $p < 0.05$.

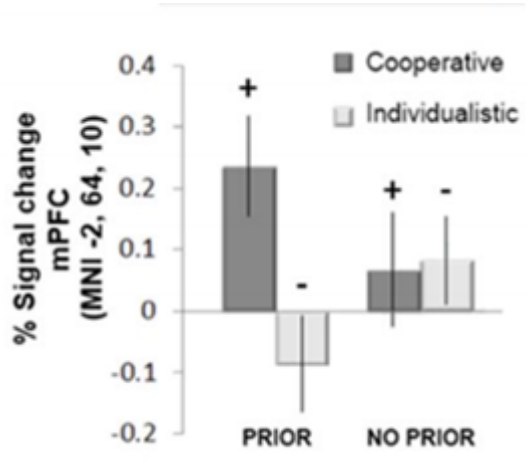


Figure 4.4.4: Functional ROI analysis in mPFC. Percent signal change by condition in the mPFC area represented in (A). Similar pattern of activity was found in the dlPFC (not reported). These regions encoded prior valence (positive and negative) that guided decision to trust at time of choice. Error bars represent SE.

4.4.3 GLM3 Effect of prior at RTG outcome

Across all RTGs, during the outcome phase of the game (see GLM2), individually estimated trial-wise prediction errors (positive and negative combined) correlated significantly with the BOLD responses in the bilateral caudate in the No Prior trials only ($p < 0.05$ FWE), (**See figure 4.4.5** and **Table 4.4.1**). On the other hand, striatal activity appeared to track estimated prediction errors in a more blunted fashion when priors were provided (**See figure 4.4.5**). Moreover, from a direct comparison between the no prior and prior conditions, we found higher activity in the left caudate for the no prior condition compare to the prior condition with a group peak MNI coordinates at -12, 20, 8 (**See figure 4.4.6**).

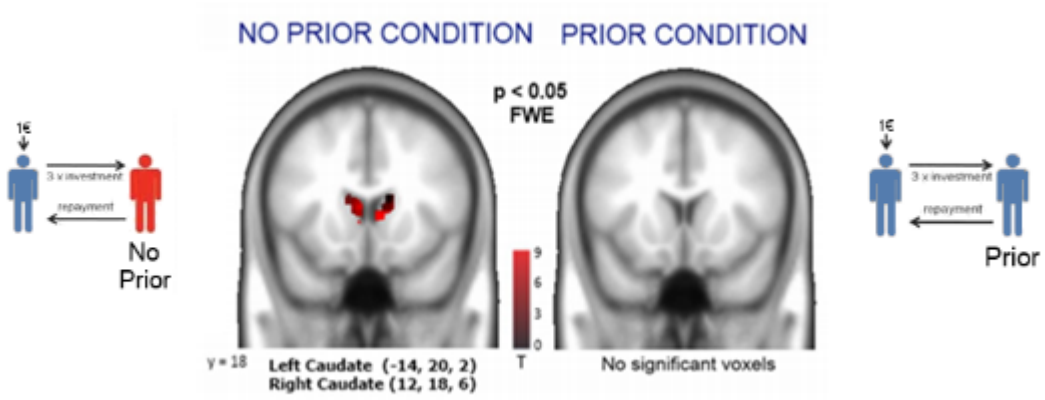


Figure 4.4.5: Brain regions parametrically correlated with the estimated Prediction Error of the best fitting RL model. Random effect fMRI analysis: Activity of the caudate showed significant correlation to the estimated PE signal in the no prior condition ($p < 0.05$ FWE cor.). Such activities were not observed in this brain area in the prior condition. Peak coordinates are given in MNI space. Colour bars indicate T-values.

4.4.4 GLM4: Individualistic versus Cooperative

Playing with a cooperative or an individualistic counterpart requested that participants adjusted their own strategy to fit the Trustees' strategy. This behaviour mechanism had a major consequence: it restricted participants to behave accordingly to the Trustee's behavior and not to their own opinion and social preference. Thus, playing with Cooperative Trustees implicated that participants' return payoffs would be higher by also cooperating and that playing with Individualistic Trustees implied to also be strategically Individualistic. However, evidence from empirical investigations has shown that most Trustors are willing to place trust

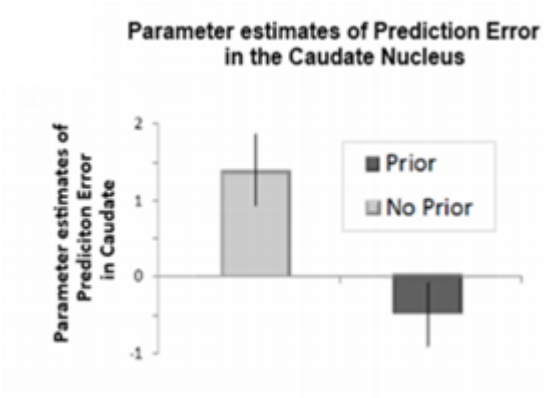


Figure 4.4.6: Parameter estimates were extracted from the left caudate (-12, 20, 8) for the direct comparison between prior and no-prior conditions. Caudate activity correlates with PE in the no-prior condition only.

and transfer their initial endowment to the Trustees in the case of repeated trust games (Berg et al., 1995; Camerer and Weigelt, 1988). Playing with individualistic counterparts as compared with cooperative counterpart was characterized by significant increases in activity of the anterior Insula, the Medial prefrontal cortex (mPFC), Anterior cingulate cortex (ACC), the orbito-frontal cortex (OFC) and the Putamen (**See figure 4.4.7**).

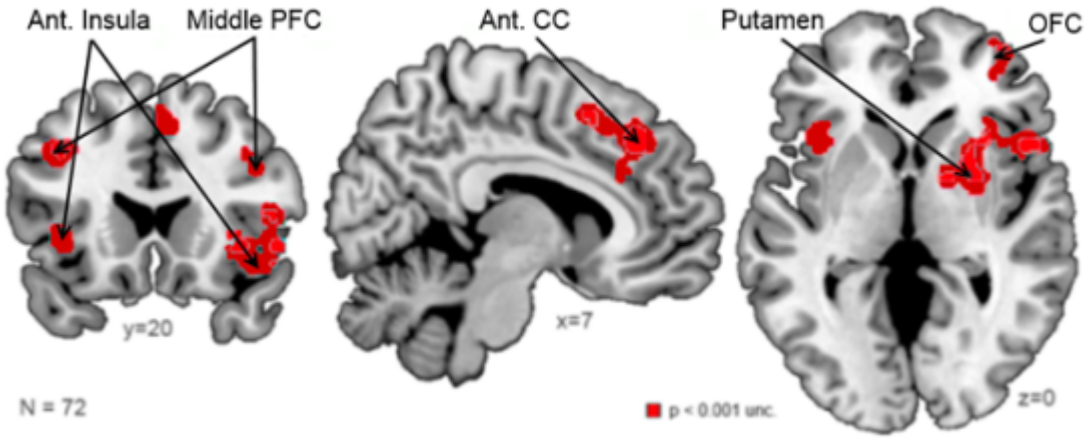


Figure 4.4.7: Differential brain activation pattern while playing with Individualistic Trustees compare to Cooperative Trustees. Illustrated are brain areas where activity exceeded threshold are $p < 0.001$ with 20 voxels extended threshold. The functional maps are superimposed on a normalized anatomical image.

4.4.5 GLM5 Violation of trust: functional connectivity analysis

Finally, we specified the changes in activity in the caudate related to violation of trust (e.g. the decisions to keep of a cooperative counterpart in response to a decision to trust of a participant) in the prior and no-prior condition (**analysis from GLM3**, **Table 4.4.1**). Results showed a stronger deactivation of the caudate in the prior condition compared to the no-prior condition (t = 6.78; **See figure 4.4.9** and **Table 4.4.1**). However, in contrast with the no-

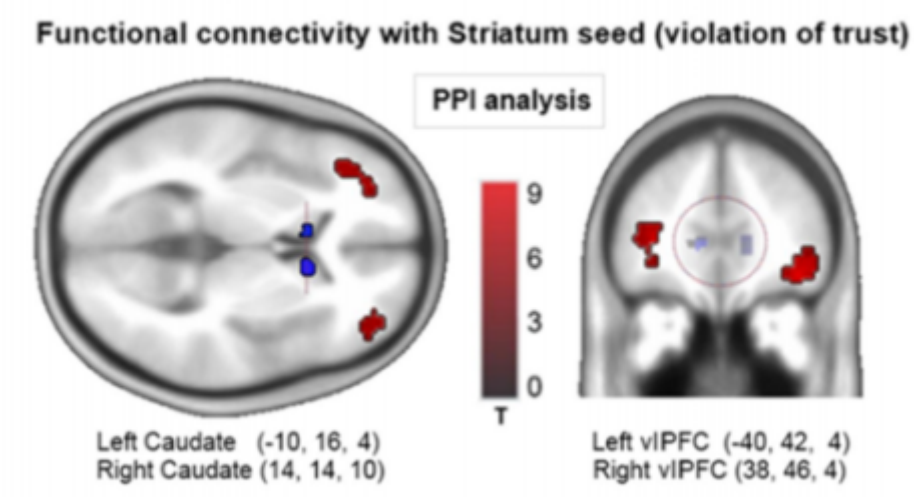


Figure 4.4.8: Functional connectivity between the caudate nucleus and vLPFC correlates with the choice to retaliate after violation of trust in the prior condition. PPI analysis. With a caudate seed, bilateral vLPFC shows stronger connectivity with this region in the prior compare to the no prior conditions.

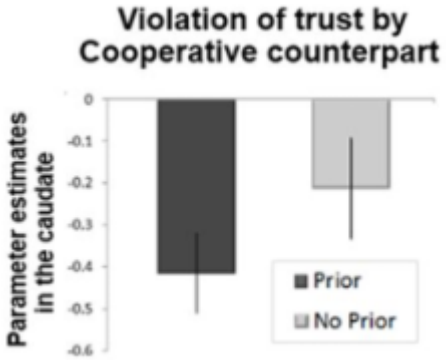


Figure 4.4.9: Reputational priors magnify striatal response to violation of trust. The caudate shows a stronger deactivation to violation of trust from a cooperative counterpart in the prior condition compared with the no-prior condition.

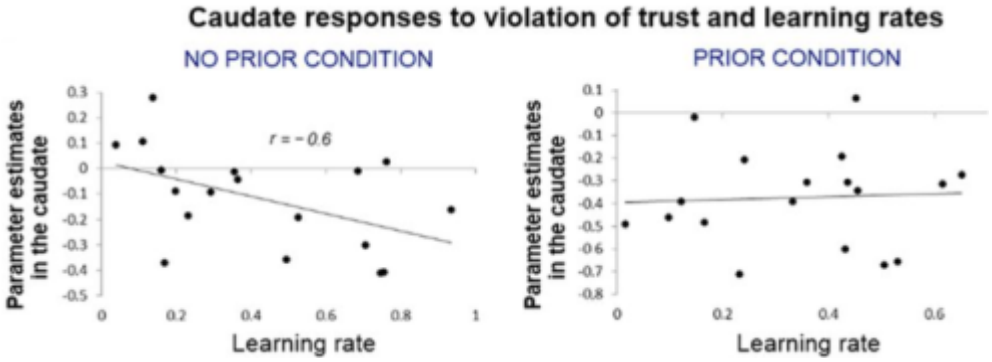


Figure 4.4.10: Striatal responses to violation of trust and learning rates. The correlation between caudate and learning rates is significant only in the no-prior condition, thus striatal responses to violation of trust in the prior condition are not reflected in learning.

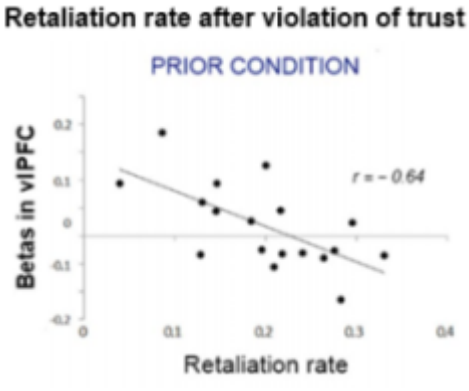


Figure 4.4.11: vLPFC prevents retaliation to violation of trust in the prior condition. vLPFC anticorrelates with retaliation rate in the Prior condition after participants experimented violation of trust from a cooperative counterpart. Spearman r = -0.6, $p < 0.009$.

prior condition, striatal deactivations to violation of trust were not reflected in behavioral responses. Indeed, the striatal activity related to violation of trust did correlate with individual learning rates only in the no-prior condition (from the model 2: $r = -0.687$, $p < 0.001$; **See figure 4.4.10**). No such correlation was found in the Prior condition. We used PPI to search for brain areas that could have mediated such striatal responses in the prior condition. We found that left and right vLPFC showed strong functional connectivity with the caudate seed region after violation of trust in the prior compared with no-prior conditions; vLPFC, left $(-40, 42, 4)$, $t = 3.73$; right $(38, 46, 4)$, $t = 6.37$, $p < 0.05$ corrected (**See figure 4.4.8**). Finally, we found that the strength of connectivity between caudate–vLPFC was anticorrelated with participants' decisions to keep following violation of trust (Spearman correlation $r = -0.67$, $p < 0.001$). Moreover, we found that the activity in the vLPFC was inversely correlated with individual retaliation rates [1] after violations of trust ($r = -0.6$, $p < 0.009$; **See figure 4.4.11**).

[1] Retaliation rates were computed as the percentage of Keep over Share choices

Analysis and locations	BA	Side	Cluster size	T	*p* value FWE cor.	MNI (mm) X	Y	Z
Prior > No Prior (GLM 1)								
Medial Prefrontal Cortex	10	—	95	8.26	6.8^{-06}	0	62	31
Ventra Tegmental Area	—	—	14	3.177	0.0032*	0	-1	-5
No Prior > Prior (GLM 1)								
Anterior Insula	44	Left	106	3.912	0.0009*	-36	-4	15
Anterior Insula	44	Right	55	3.450	0.0017*	38	3	10
Parametric regression: Choice (GLM 2)								
Medial Prefrontal Cortex	10	—	87	6.562	2.7^{-06}	-2	64	10
Lateral Prefrontal Cortex	46	Left	122	5.987	7.8^{-05}	-38	38	32
Lateral Prefrontal Cortex	46	Right	109	6.342	2.1^{-06}	30	38	34
Superior Parietal Lobule	48	Left	77	5.01	6.7^{-04}	-38	6	24
Parametric regression: Outcome No Prior condition (GLM 2)								
Caudate Nucleus	—	Left	77	7.091	8.9^{-06}	-14	20	2
Caudate Nucleus	—	Right	56	8.298	7.9^{-06}	12	16	8
Violation of trust in the Prior condition (GLM 3)								
Caudate Nucleus	—	Left	82	6.78	2.8^{-06}	-10	18	11
Caudate Nucleus	—	Right	56	6.34	2.4^{-06}	12	21	5
Cooperative > Individualistic (GLM 4)								
Lateral Middle Prefrontal Cortex	48	Left	49	4.687	3.2^{-05}	-39	19	22
Lateral Middle Prefrontal Cortex	48	Right	72	4.618	4.78^{-05}	40	22	24
Anterior Middle Prefrontal Cortex	46	Right	57	4.259	7.3^{-05}	-32	36	26
Insula	48	Left	85	3.750	0.0001*	-38	14	-2
Insula	48	Right	122	3.644	0.0002*	57	17	-6
Anterior Cingulate Cortex	32	—	151	3.534	0.0003*	5	30	27
Putamen	—	Right	98	3,518	0.0012*	25	9	4
Middle Orbito Frontal Cortex	46	Right	59	3.525	0.0003*	39	51	-2

Table 4.4.1: Activations correlated with contrasts of interest Note: BA, Brodmann area; * uncorrected statistics.

4.5 Discussion and Conclusions

Reputation-based social decision-making has been investigated both by theoretical and empirical studies (Boero, Bravo, Castellani, and Squazzoni, 2009; Camerer and Weigelt, 1988; Fudenberg et al., 1990), however research on its neurocognitive bases is still in its infancy. Though it is rather unlikely that, in daily decisions, people possess absolutely no prior neither contextual information on who they interact with, the growing literature using RTGs in fMRI focused mainly on situations in which strictly no priors are available (McCabe et al., 2001; King-Casas et al., 2005; Krueger et al., 2007). Only two recent fMRI studies investigated how social priors (i.e. the moral character of their counterparts) affect the way people engage in RTGs (Delgado et al., 2005; Dominic S Fareri et al., 2012). These studies however did not completely isolate the effect of priors on trust (prior-based trust) by confronting them to identical conditions with no priors (interaction-based trust). Our experimental setting is the first to allow this direct comparison. The main goal of our study was to determine whether, and how, reliable reputational priors affect initial decisions and subsequent learning mechanisms at both the behavioral and neural level. From a behavioral point of view, we show that priors affect decisions to trust in at least 2 ways: 1) in initial stages of the interaction, participants clearly chose to trust or distrust according to the positive or negative reputation of their counterparts; furthermore, 2) players tend to keep relying on reputation priors, even when their counterpart's behavior was inconsistent with it. As a consequence, and since priors were accurate predictors of trustworthiness in our study, players earned more when reputational cues were available than when they were not.

4.5.1 mPFC encodes reputational priors

From a neural point of view, our fMRI results revealed that the presentation of a new counterpart yielded enhanced activation in the mPFC when accompanied by a prior (irrespective of it signaling a positive or negative reputation). We suggest that the enhanced mPFC activity may reflect the fact the prior information reduced the uncertainty about the behavior of the other faced by participants when beginning a new RTG. Indeed, this region has been previously implicated in uncertainty resolution in interactive contexts (Yoshida and Ishii, 2006). This is furthermore consistent with the inverse activation pattern observed in the insula, which showed stronger activity when priors were not available, consistently with previous findings reporting a role for this region in tracking increased uncertainty (Preuschoff, Quartz, and Bossaerts, 2008).

4.5.2 mPFC and dLPFC encode the value of reputation priors

At time of choice, the valence of priors elicited dissociable activation patterns when integrated with the behaviorally estimated (from the prior-based RL model) option values (Qt). Specifically, the mPFC and dLPFC differentially responded to cooperative versus individualistic counterparts, however, only when priors were available. As reported in previous studies, our results suggest that this brain network keeps track of contextually modulated decision values over trials, and doing so improves participants' performance (Wunderlich et al., 2009).

As reputational priors conveyed information on the social attitudes of counterparts in our study, this activation is also consistent with a well established role of the mPFC in ascribing attitudes to others (**?**), and anticipating their choices (Krueger et al., 2007; Hampton et al., 2008; **?**). Thus the mPFC is encoding a first response to reputational priors as well as the effect of priors during subsequent interactions. This is in accordance with findings from humans (Rilling et al., 2002; Hampton et al., 2008) and nonhuman primates (Barraclough et al., 2004) on the role of the PFC in encoding value-related signals in repeated interactions.

4.5.3 Caudate nucleus encodes reward PE only when prior information is not provided

Consistent with previous studies, trial-by-trial prediction errors estimated by RL models correlated with activity in the striatum (Bunge, 2004; McClure, Laibson, Loewenstein, and Cohen, 2004; O'Doherty, Dayan, Schultz, Deichmann, Friston, and Dolan, 2004; King-Casas et al., 2005; Schönberg et al., 2007) but, critically, only when no priors were available. This confirms a role for the caudate in tracking the difference between expected and obtained outcomes in RTGs, triggering learning. However, when priors were available they appeared to prevent participants from reinforcement-based learning, which was reflected in the reduced covariance between caudate responses and estimated prediction errors.

4.5.4 Priors magnify reward-prediction error signals in the caudate nucleus

As regards the striatal activation patterns, these are well-aligned with an established role of the striatum in tracking reward contingencies, in both non-social (O'Doherty et al., 2004) and social domains (Delgado et al., 2005; King-Casas et al., 2005; Jones et al., 2011). More specifically, the observed patterns are consistent with the idea that the caudate mediates the neural computation of reward prediction error (RPE). Indeed, we observed RPE-pliant signals in the caudate only when no priors were provided, while the same signals appeared blunted when

priors were available. Previous studies on non-social tasks (Li et al., 2011; Doll et al., 2009, 2011) and social tasks (Delgado et al., 2005; Biele et al., 2011; Fareri, Chang, and Delgado, 2012) have shown that, when priors are available, participants tended to hinge on to them, and to relatively discount the impact of the outcomes of their past decisions.

However, in addition to the previous studies, our results show that the presence of priors magnifies striatal deactivation to violations of trust (i.e. when a counterpart with positive reputation, as opposed to no reputation, violated trust), rather than blunting their response. Why previous studies didn't find such magnified response due to violation of priors requires further investigation, though several hypotheses are possible. For instance, two studies (Delgado et al., 2005; Fareri et al., 2012) focused on the subset of unreliable priors, that is, on priors that carried no information on trustees' actual choices; it is likely that, in such a scenario, participants were gradually learning to disregard such priors, converging towards their extinction rather than exploitation. On the other hand, the opposite may have occurred in a more recent study on the non-social domain (Li et al., 2011), in which priors were perhaps too reliable. Indeed, in that study, agents were explicitly instructed on the precise probabilities of outcomes, which may have reduced their surprise when infrequent, though anticipated losses occurred. In both these previous studies, the space for learning via priors may have been reduced, as the actual prior-to-reward contingencies appeared either non-existent (Delgado et al., 2005; Fareri et al., 2012), or already completely exploited (Li et al., 2011). It is also possible that the different methods used to instil priors tapped on different neural mechanisms: Delgado and colleagues (2005) provided short descriptions of the "moral character" of counterparts, whereas Fareri and colleagues (2012) used direct evidence from previous experience (i.e. playing a ball task). Such methods of instilling priors may have also made them more salient or intuitive and, as a result, harder to extinguish in spite of conflicting evidence. On the other hand, our task reported on characteristics of counterparts that were possibly more directly linked to the main task (i.e. the priors were based on results indicating the extent to which one cares about his own payoffs and that of others - SVO task). Further investigation specifically manipulating prior reliability should clarify some of the points of divergence. Until then, the open question in our study regarded the reason as to why striatal deactivations to trust violations were not leading to behavioral adjustments when priors were available.

4.5.5 VLPFC Caudate stronger functional connectivity preventing retaliation

On the other hand, when priors were present, we suggest that the impact on learning of the striatal deactivations to violations of trust may have been disrupted by other brain areas. Our

results are in line with attributing this role to the vLPFC, which we found to functionally correlate with such striatal deactivations. In particular, the strength of connectivity between caudate and vLPFC was stronger in the prior compared to the no prior condition. We thus propose that the vLPFC contributes in maintaining choices aligned with the reliable prior beliefs, when beliefs momentarily conflict with observations. This might occur by compensating for the relatively automatic behavioral changes to reward prediction error signals. In line with this interpretation previous literature has implicated the vLPFC in top-down cognitive control by biasing processing in other brain regions towards contextually appropriate representations (Cohen, Mcclelland, and Dunbar, 1990; Miller and Cohen, 2001). Furthermore, not only the vLPFC plays a role in modulating bottom-up fashion cognition processes, but this area has also been found to play a role in goal-directed behavior (Souza, Donohue, and Bunge, 2009; Valentin et al., 2007). In conclusion, our study integrates theories and methods from cognitive neuroscience, economics, and RL to gain a greater understanding of how reputation priors are encoded in the brain and how they affect learning to trust anonymous others. Our findings suggest that priors influence both initial decisions to trust and the following learning mechanisms involved in repeated interactions. Specifically, the present study showed that reputational priors magnify striatal responses to violations of trust. However, when such priors are reliable, other phylogenetically younger brain regions involved in higher cognition may contribute to keep decisions anchored to those priors, thus relatively discounting the weight of conflicting evidence. The interplay between striatum and lateral orbitofrontal cortex may prevent unnecessary retaliation when others violate our trust, and thus constitute an important neuro-cognive mechanism that favors social stability.

All trust involves vulnerability and risk, and nothing would count as trust if there were no possibility of betrayal.

Robert C. Solomon

5

Simultaneous eye-tracking and galvanic skin response

5.1 Background

In the previous chapters, we have shown the behavioral, computational and neural patterns underlying adaptive social learning and how reputational priors affect this learning mechanism. We investigated two mains effects: the absence or presence of reputation information and two probabilities of reciprocation (high and low). In the first study, the responses of participants in long RTG (10 exchanges) were best predicted by an Adaptive-Belief RL model which captures long-term relationships between reputation and experienced trustworthiness. However, when interactions end stochastically (RTGs are shorts), people rely more on prior expectations driven by reputations and this is best explained by a Prior-Expectation RL model. Indeed, the first trial of the interaction with a new counterpart provides the best behavioral discrimination between the prior and no prior condition, and this is reflected in enhanced Medial-Prefrontal cortex activity. This motivated us to further explore how reputation priors affect first trustworthiness judgments in single-shot Trust Games.

Moreover, the two first studies suggest that participants are particularly sensible to trust violation when partners have high pro-social reputation. Trust violations happen when the trustors (e.g. our participants) have trustworthy expectations of the trustees and when these later contradict expectations. Previous studies report that trust violation can cause severe psychological and relationship damages (Slovic, 1993; Lewicki, McAllister, and Bies, 1998), and (Chapter 7 in Roy J. Lewicki, 1996) and even in some cases, lost trust can never be restored. In their studies, Lewicki and colleagues propose theoretical models for trust violation in which they assume that trust betrayal may permanently harm trust. However, no study has investigated on the memory, emotional and psychophysiological consequences of violating trust.

In this new study, we decided to focus on single-stage Trust Games when no learning takes place. Participants played as trustor in single TGs with partners of different pro-social nature with or without reliable reputational priors. We used probabilistic responses to simulate partners with different reputations (from competitive to altruistic) while recording electrodermal activity as an autonomic index of affective state as well as eye movements.

Electrodermal responses and eye-movements were recorded while participants played single-stage TGs with different anonymous partners. We used eye-tracking technique to investigate whether reputational priors affected eye saccades when viewing face-stimuli of their counterparts. Extensive research on face perception has shown that people use holistic and features-based processing to evaluate facial expression when they encounter new partners (Dalton, Nacewicz, Johnstone, Schaefer, Gernsbacher, Goldsmith, Alexander, and Davidson, 2005; Stacey, Walker, and Underwood, 2005; Belle, Ramon, Lefèvre, and Rossion, 2010; Eisenbarth and Alpers, 2011). Critical information in inter-personal communication is evaluated from the eyes and, to a reduced extent, the nose and the mouth. In addition, these evaluations have been found to be predictive of people's ability to perceive trustworthiness and aggressiveness (Bar, Neta, and Linz, 2006; Willis and Todorov, 2006). In one recent study, using oxytocine (OT), authors have shown that that OT nasal administration (enhancing trust) modified the way people evaluate faces of others, increase the amount of eye-gaze (Guastella, Mitchell, and Dadds, 2008). In this new study, we tested the hypothesis that reputational prior reduce uncertainty about the counterpart's type and thus decreases the amount of eye-exploration. It was hypothesized that participants' saccades when scanning face-stimuli of counterparts who were associated with a reliable prior would be less important than when they were exploring the face-stimuli of counterparts with no reputation.

Recording electrodermal activity is a well-established method in psychophysiology that informs about the activity of the autonomic nervous system. Its relationship with emotional arousal and emotions has been proven to be robust in a lot of experimental studies. In a recent neuroeconomics study, electrodermal responses were found to be higher for unfair offers

compare to fair offers and were concomitant with the rejection of unfair offers when participants played the Ultimatum Game (Dunn, Evans, Makarova, White, and Clark, 2012). We speculated that decision making with reliable reputation information may decrease electrodermal activity responses during a trust-related social situation since uncertainty is reduced. We also hypothesised to find interactions between trust and reputation at behavioral and psychophysiological level. Additionally, we postulated that violation of trust of counterparts with the highest pro-social reputations would be directly linked to physiological changes (i.e. autonomic nervous system responses at time of violation of trust). In line with the last hypothesis, we investigated whether intra-individual differences in participants' electrodermal responses (EDRs) recorded when observing violation of trust predicted subsequent recognition of partners who had betrayed trust.

5.2 Experimental design and methods

5.2.1 Participants

Forty-four Italian college students of the University of Trento within the ages of 19 to 28 ($M = 22 \pm 2$) were recruited through online announcements. There were 11 males and 33 females and 43 of the 44 participants were right-handed. Participants had normal or corrected-to-normal vision, no history of psychiatric, neurological or major medical problems and were not under psychoactive medications at the time of the study. They gave written informed consent to participate in an eye-tracking and electrodermal activity decision making experiment and decided to either be paid 10 euros. In addition to this, the participant received an additional payment based on their performance in the two tasks The study was approved by the local institutional ethical board of the University of Trento. Because of technical issues with the eye-tracking, only 31 participants remained in the analysis of the eye-mouvements.

5.2.2 Task

In the experiment, participants played one trial of the Trust Game (TG) (Berg et al., 1995) with several counterparts (i.e. each trial with a different counterpart). In the game, the participant, also known as "Investor", is given 1 euro and has to decide whether to share half of his endowment with an anonymous counterpart ("Trustee") or keep the money. If the participant decides to share, the amount that the Trustee receives is tripled by the experimenter and added to the amount he currently has (1+3 euros = 4 euros) and the Trustee decides whether to share or keep the money with the Investor. If he decides to share the money, the Investor

and the Trustee end the game with 2 euros each, but if the Trustee decides to keep the 4 euros, the Investor receives nothing. Each trial begins with a fixation cross for 1000 ms, followed by the presentation of a prior with or without an occluder covering the prior, which is then followed by a picture of the face of the counterpart (**See figure 5.2.1**).

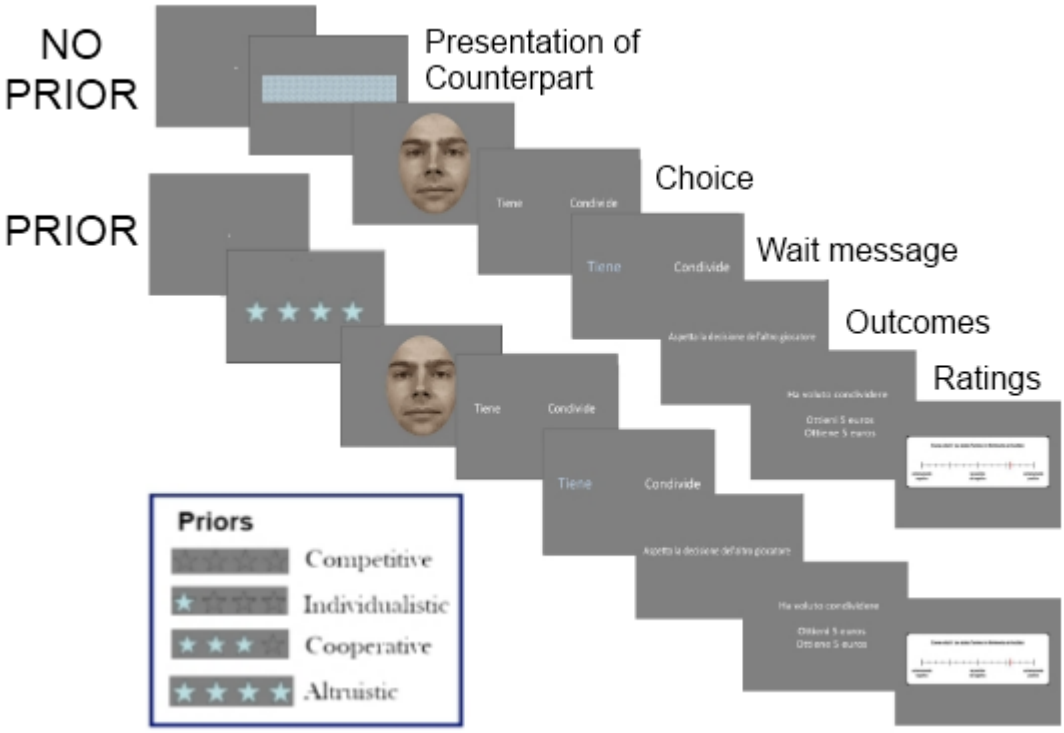


Figure 5.2.1: Experimental design. One round of the two-player repeated trust game (RTG). P1 is the payoff of the participant, who always plays as investor; P2 is the payoff of his counterpart, who plays as trustee. Before each round both players are endowed with 1 euro. The participant moves first and chooses either to "keep" or "share" If he keeps, both players maintain their initial endowments. If he shares the participant"s endowment is multiplied by 3 and passed to the counterpart. The trustee then decides whether to share in turn (by returning 2 euros), or to keep (by returning nothing). RTGs consisted of several consecutive rounds with a same counterpart. Participants played with many different counterparts and were told that their counterparts had already made their choices.

Faces of Caucasian males, aged 20-60 years old ($M = 34.09, SE \pm 11.19$) were selected and extracted with permission from the FERET database of colored facial images (Phillips et al., 2000). Two databases of images were created, 64 were used for the experiment while an additional 32 were used for the post-experiment questionnaire. All 96 images were controlled for attractiveness, emotion and race. After the main experiment, participants took a five-minute break before proceeding to the post-experiment questionnaire - the memory task. The purpose of the memory task is to ascertain whether the outcome of the trust game and satisfaction regarding the interaction influences the recognition of the Trustee's face. Participants were shown a face of either a previous counterpart in the experiment or a face not previously

presented. The question *"Do you remember this person?"* was presented and they were asked to indicate whether they remembered the face or not by pressing on the left (Yes) or right (No) buttons of the mouse.

5.2.3 Experimental conditions

Trustees were divided into four types (Altruistic, Cooperative, Individualistic and Competitive), pertaining to their probability of reciprocation, or the probability of choosing to Share their money with the Investor. Altruistic trustees reciprocate 90% of the time, Cooperative trustees reciprocate 70% of the time, Individualistic trustees reciprocate 30% of the time and Competitive trustees reciprocate 10% of the time.

The presence and absence of a reputation prior was also manipulated for this study. Half of the trustees in each type were presented with a reputation prior pertaining to their type, while the other half came with no prior information. Participants were informed that all trustees completed the Social Value Orientation (SVO) questionnaire (Van Lange, 1999) and that they may or may not be provided with information about the trustees' performance in the questionnaire. For the prior condition, stars were presented to show the type of trustee based on the questionnaire. Altruistic trustees had 4 stars, Cooperative trustees had 3 stars, Individualistic trustees had 1 star while Competitive trustees had 0 star ratings. For the no prior condition, participants were presented with an occluder to cover the star ratings.

Once the trust game with the counterpart is finished, the participant was asked to give a satisfaction rating of his previous interaction with the Trustee on a scale of -50 to +50.

Participants were paid a fixed amount of 10 euros for participation and additional money was paid based on their performance in the two tasks. For the main experiment, participants were paid an additional 0, 5 or 10 euros based on the outcome of a randomly extracted trial in the trust game. Furthermore, participants were paid up to 5 euros more based on their accuracy rates in the memory task of the post-experiment questionnaire (50% or less accuracy = 0 euros, 51-60% = 1 euro, 61-70% = 2 euros, 70-80% = 3 euros, 80-90% = 4 euros, 90-100% = 5 euros). Hence, participants earned a total of 10-25 euros ($M = 18.5 \pm 3.69$).

5.2.4 Task and electrodermal procedures

Participants were first given oral instructions about the task in Italian. After being debriefed about the experiment task, electrodes from the Biopac MP150 were placed in two of the participant's fingers in the non-dominant hand (usually the left) and he or she was fitted with

the eye-tracking helmet (**See next paragraph**). The electrodermal response was then amplified via the amplifier module GSR100C. Participants were instructed to follow the dots with their eyes. Participants did 5 practice trials and then 64 experimental trials (8 per condition: Prior Altruistic, Prior Cooperative, Prior Individualistic, Prior Competitive, NoPrior Altruistic, NoPrior Cooperative, NoPrior Individualistic, NoPrior Competitive) with a short self-determined resting period provided after 32 trials. **Figure 5.2.1** shows the timeline of one trial. They were given a short rest period followed by debriefing about the post-experiment memory task, and 3 practice trials were presented before completing 64 trials in the memory task. Eye-tracking and electrodermal activitywere measured in both the experiment and post-experiment trials.

5.2.5 Eye-tracking procedures

Participants were seated in a chair with a soft head restraint to ensure a viewing distance of 60 cm to the monitor. Presentation of the stimuli was performed using a custom made program written using the Matlab Psychophysical toolbox. Eye movements were monitored and recorded using an Eyelink II system (SR. Research Ontario Canada) with a sampling rate of 500 Hz. A fixation was defined as an interval in which gaze was focused within 1° of visual angle for at least 100 ms (Manor and Gordon, 2003). A nine-point calibration was performed at the beginning of each block. Calibration phase was repeated until the difference between the different positions of the points on the screen and the corresponding eye locations was less than 1°. After the calibration phase, a nine-point validation phase was performed (similarly to the calibration phase) to make sure that the calibration was accurate. Recalibrations were performed if needed, and eye-tracking was stopped if these were unsuccessful. Before the beginning of each trial a drift correction was performed (except for the first trial of each block). Then, a fixation point was presented in the same position of the last point of the drift correction (last point of the validation phase for the first trial of each block) for 500 ms. The fixation point was located in the middle of the two possible choices (Keep or Share). To minimize biases related to the starting fixation point and anchoring effects, the position of the words (Keep or Share) were counterbalanced. Eye movements were recorded during the entire time of the trial. To minimize noise, information displayed on the monitor was limited to few words and numbers. In order to calibrate the areas of interest of each face, the pictures were spatially transformed (using Photoshop) so that the center of each eye (the iris) and the center of the mouth (where the lips meet) correspond to the same coordinate positions in the transformed picture.

5.3 Data Analysis

5.3.1 Behavioral Data

In order to look at whether prior information has an effect on the decision to trust, repeated measures ANOVA was performed with Prior information (No prior, 0-star prior, 1-star prior, 3-stars prior, 4-stars prior) as independent variable and percentage of decisions to "share" per participant as dependent variable. Repeated Measures ANOVA was also performed on reaction times at time of choice, in order to determine if lesser ambiguity provided by prior information lead to faster reaction times. The outcome of a previous interaction with a counterpart is also hypothesized to influence the decision to share with a similar counterpart. Hence, logistic regression was performed in order to determine if the outcome of the previous interaction with a counterpart type could predict the participants' decision to share for a subsequent counterpart of the same type. It is hypothesized that there would be a higher negative emotional valence when the outcome of the interaction is negative (participant shared and counterpart kept), and the counterpart does not act in congruence with the prior information. Hence, the emotion scale was analysed through Repeated Measures ANOVA, extracting only the trials with Prior condition and those when the participant shared and the opponent kept. For the post-experiment questionnaire, it is hypothesized that participants would have a higher accuracy for the faces of counterparts whom they decided to trust (participant shared), and for faces of counterparts who betrayed them (participant shared and counterpart kept). Hence, paired t-tests would be performed on accuracy rates, comparing counterparts for whom participants shared vs. kept, and where participants shared and counterpart kept vs. counterpart shared.

5.3.2 Electrodermal preprocessing

In order to preprocess the galvanic data, we used the toolbox developed by Mateus Joffily (Toolbox for Electrodermal Activity (EDA) analysis) based on Matlab. The first step of our analysis was to verify the quality of the data acquired during the experimental session by generating and visually examining plots of the electrodermal signal at different steps of the preprocessing (**See raw data example - figure 5.4.1**). Secondly, we applied both a high-pass filter in order to remove most of the tonic changes as well as the slow drifts (Freedman, Scerbo, Dawson, Raine, McClure, and Venables, 1994), and a low-pass filter to remove the high-frequency noise (5th-order low-pass Butterworth filter with cutoff frequency at 1Hz). The resulting band-pass filtered signal was then downsampled respecting the Shannon-Nyquist sampling

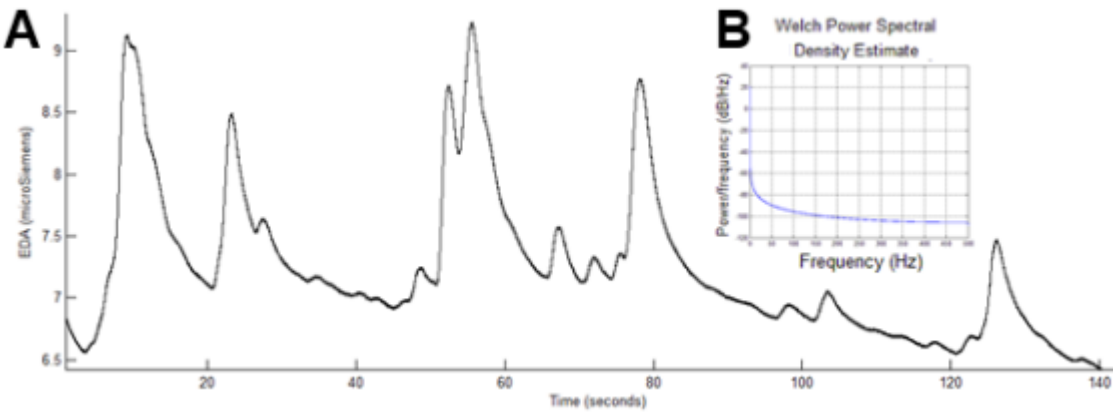


Figure 5.3.1: A. Example of raw galvanic skin response for one participant over time. B. Power Spectrum representing the frequency of the signal.

rule (i.e. the sampling rate fs must be greater than twice the bandwidth of the signal). Once the data was preprocessed, we ran an automatic detection algorithm that finds Electrodermal Response (EDR) in the galvanic responses by calculating a first time-derivative of the data using a difference function, and detecting the main characteristic of the galvanic response (the valleys and peaks) when the derivative changes sign. Onsets of valleys are identified through a negative to positive zero crossing, whereas peaks are defined by an opposite change (positive to negative zero crossing). The amplitude of the galvanic response is computed as the difference between the amplitudes at the peak and valley levels. The slope corresponds to the rate of changes and it is computed as the ratio between the 2 EDR amplitudes and the ride time. Overlapping EDRs were disjointed at minima in the first derivative and used as separate

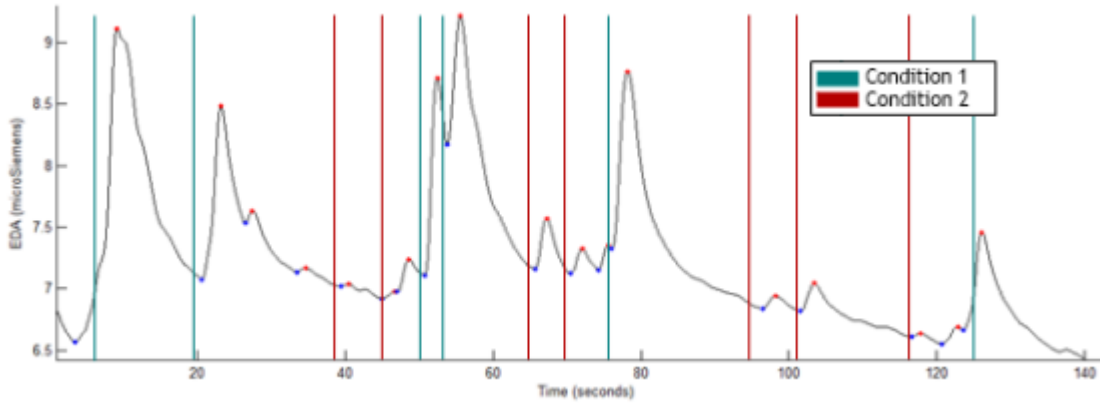


Figure 5.3.2: Representative regressor for one participant distinguishing between 2 conditions above preprocessed EDA. Note that the red dots represent the peak estimates and the blue dots represent the valleys.

EDRs. Short responses (below 0.005 μS) were excluded from the analysis since they are likely to indicate environmental noise. In order to correct for slope, we log-transformed the EDRs,

and range-corrected the data using each participant's maximum range response. Thus, we report all our analysis in units of range-corrected logE(μ.S) (**See preprocessed data example - figure 5.4.2**).

5.3.3 Electrodermal analysis

The main goal of this study was to assess the difference between *"direct-based"* interaction and *"reputation-based"* interaction; if the strength of reputational priors affect *"reputation-based"* interaction and if the link between the strength of the reputational prior in the experiment will facilitate recall of faces in the post-experiment task. Accordingly, the following analyses investigated two main onsets of time: at time of decision-making and at time of outcomes.

1. We first applied a search grid algorithm with a moving window EDR average that allows us to determine the onsets for choice and outcome periods at 1.25 seconds and 1.4 seconds respectively. These two onsets are the time points after participants were asked to decide whether to trust the counterpart or not, and when they saw the outcome of the game. The EDR magnitudes that were used in our statistical model correspond to individual mean responses across all trials from all onsets corresponding to a specific condition. In order to verify that our data were not contaminated by habituation effects, we looked for individual mean responses in both the experiment and the post experiment task, and tested for habituation effects.

2. Our first analysis examined whether the skin conductance level at time of choice was sensitive to decision-making and outcome anticipation. We probed whether this activity was larger for the decision to share than for the decision to keep and how these decisions were influenced by prior conditions.

3. Secondly, we looked at the EDR at time of outcome and investigated the discrepancy between observed outcomes and anticipated outcomes. Additionally, we checked for correlation between the reported emotional levels and the EDR recorded.

4. Then we analyzed participants' EDRs when evaluating the outcomes of decisions in the experiment. We correlated individual changes in EDR magnitudes with individual changes in the percentage of correct recognition in the post-experiment questionnaire. We correlated participant's EDR magnitudes and their mean emotional rating when observing the outcomes of each interaction. To assess whether there is a differential response magnitude to the experience of betrayal and trustworthiness in the prior and the no prior conditions, we calculated a Violation-Trust-EDR score for each participant, and correlated it with the correct recall in the second session. This score was the

absolute differential EDR magnitudes at time of outcome in two situations: (1) while being betrayed by a presumably cooperative counterpart, and (2) while being repaid for placing trust with the same type of counterpart: "Violation-Trust-EDR". To assess whether the presence of a prior information for "Violation-Trust-EDR" conditions and the differential scores contribute to the variance in recognition in the memory task, we used a hierarchical multiple regression analysis using "Violation-Trust-EDR -Prior", "Violation-Trust-EDR -No-Prior" and the differential scores as regressors. In order to discern the relationship between violation of trust and percentage of recognition, regardless of the effect of prior information, we used force entry for "Violation-Trust-EDR-No-Prior" in the first step. Consecutively, we used the two variables of "Violation-Trust-EDR-Prior" and "Violation-Trust-EDR-No-Prior" scores to investigate whether they captured some of the variance. In order to evaluate regressors in the model without accounting for their order of entry, we used a stepwise procedure. For each variable, we identified outliers as responses that were over 3 standard deviations from the mean, and Pearson correlations that were significant at $p < 0.05$, one-tailed, unless otherwise specified.

5.3.4 Eye-tracking analysis

For each pictures, we define 4 areas of interest (facial-AOIs, as seen in the **figure 5.4.3**), one for each eye, the nose and the mouth. The 2 eyes-AOIs had a circular shape with 11182 pixels; the nose-AOI had a circular shape with 10073 pixels and the mouth-AOI an ellipse with 12302 pixels. We centered the eyes-AOIs in the iris. This allowed us to avoid the possibility that small errors in the calibration procedures could result in a wrong allocation of the eye-tracking parameters. This is especially useful for those situations in which the parameter was located on the border of a facial-AOI. Facial-AOIs do not cover the face area entirely but just the part of the face known to be relevant for facials extractions and never overlap. In this way, facial-AOIs include only saccades whose interpretations are not ambiguous. Because I report just preliminary results, only one type of variable is described, corresponding to the number and type of saccades participants made when looking at the face of others. These saccades relate to the eye gaze transitions made by the participants from one region of the screen to another. For the analysis, we distinguished between (1) the facial-relevant saccades being the transitions from one facial-AOI to another facial-AOI and (2) the non-relevant saccades as being the transitions starting or ending outside and facial-AOI.

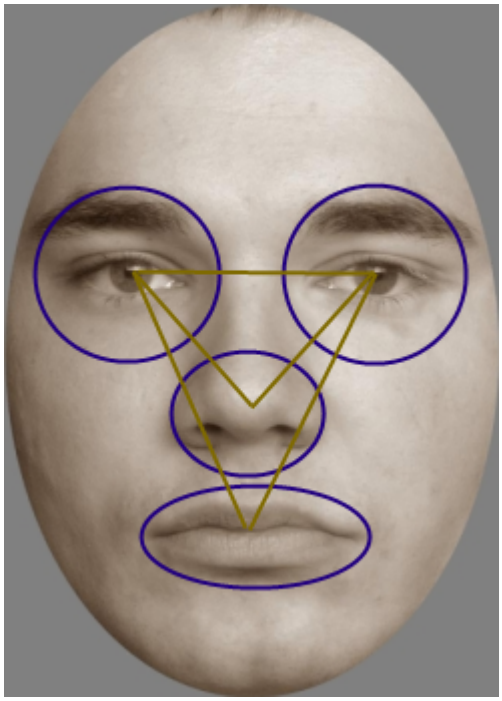


Figure 5.3.3: Areas of interest. Example of areas of interest (AOIs) for one face-stimulus. For each face stimulus, we defined AOIs: the eyes, nose, and mouth. The brown lines represent examples of the relevant saccades for facial perception.

5.4 Results

5.4.1 Effects of prior manipulation

In order to check for the prior effect in the direct-based and reputation-based conditions, we extracted EDRs at time of choice in situations when participants had a prior versus when they didn't. We also examined conditions when participants decided to share in the prior condition as compared with the no prior condition. First, we found that EDRs were significantly stronger when participants shared ($M = 0.748$, $SE \pm 0.3$), as compared to kept ($M = 0.415$, $SE \pm 0.09$), $t(41) = 5.05$, $p<0.001$, $\eta^2 = 0.63$ (**See figure 5.5.1 A**). EDRs were also found to be significantly stronger for choices made in the no prior condition ($M = 0.975$, $SE \pm 0.07$) as compared to the prior condition ($M = 0.525$, $SE \pm 0.02$), $t(41) = 3.74$, $p = 0.001$, $\eta^2 = 0.45$ (**See figure 5.5.1 B**). For counterparts that shared, we found higher mean EDRs for counterparts that were presented without a prior ($M = 1.123$, $SE \pm 0.45$) than with a prior ($M = 0.6$, $SE \pm 0.43$, $t(41) = 3.32$, $p<0.001$, $\eta^2 = 0.41$). In addition, mean EDRs for counterparts that kept were lower for those presented with a prior ($M = 0.259$, $SE = 0.39$) than without a prior ($M = 0.681$, $SE \pm 2.85$), $t(41) = 2.78$, $p<0.006$, $\eta^2 = 0.29$) (**See figure 5.5.2**).

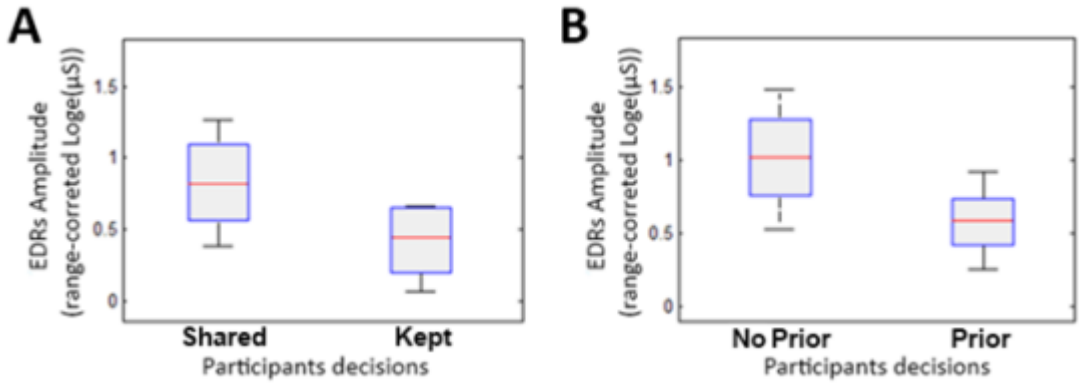


Figure 5.4.1: A. Main effect of decision of the participant at time of choice. We found an overall effect of the type of decision. The EDR was higher for decisions to share than for decision to keep. B. We found the EDR was higher for decisions made in the no prior condition than for decisions made in the prior condition.

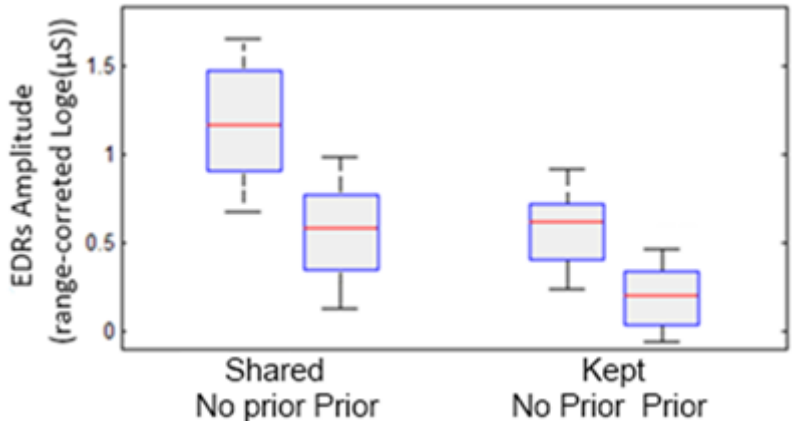


Figure 5.4.2: Differential effect of decision of the participant at time of choice in the prior and no prior conditions.

The results of the ANOVAs with the factor (prior/no prior) and time (first half/second half of the experiment), that were computed separately for conditions in which participants decided to share and keep in the TG, showed no difference overtime. This shows that EDRs were not affected by habituation (**See Table 5.5.1**).

5.4.2 Effect of inconsistent outcomes and recognition

On average, people recall the correct picture in 73.65% (SE = 5.5) of the cases. They recognize correctly 76.6% of the pictures in the no prior condition and 70.74% in the no prior condition (SE = 6.3), but this trend was not significant (p=0.23). Analysis comparing EDR responses in the experiment and post-experiment questionnaire indicate no effects of time, indicating no

ANOVAs			
	F(1,41)	P	η^2
Factor decision - Shared	28.39	<0.001	0.41
Factor decision - Kept	24.91	<0.001	0.37
Factor time - Shared	0.051	0.84	0.002
Factor time - Kept	0.076	0.82	0.003
Interaction decision x time - Shared	0.65	0.44	0.033
Interaction decision x time - Kept	0.16	0.69	0.008

Table 5.4.1: Anovas. Results of ANOVAs with the factor (prior/no prior) and time (first half/second half of the experiment).

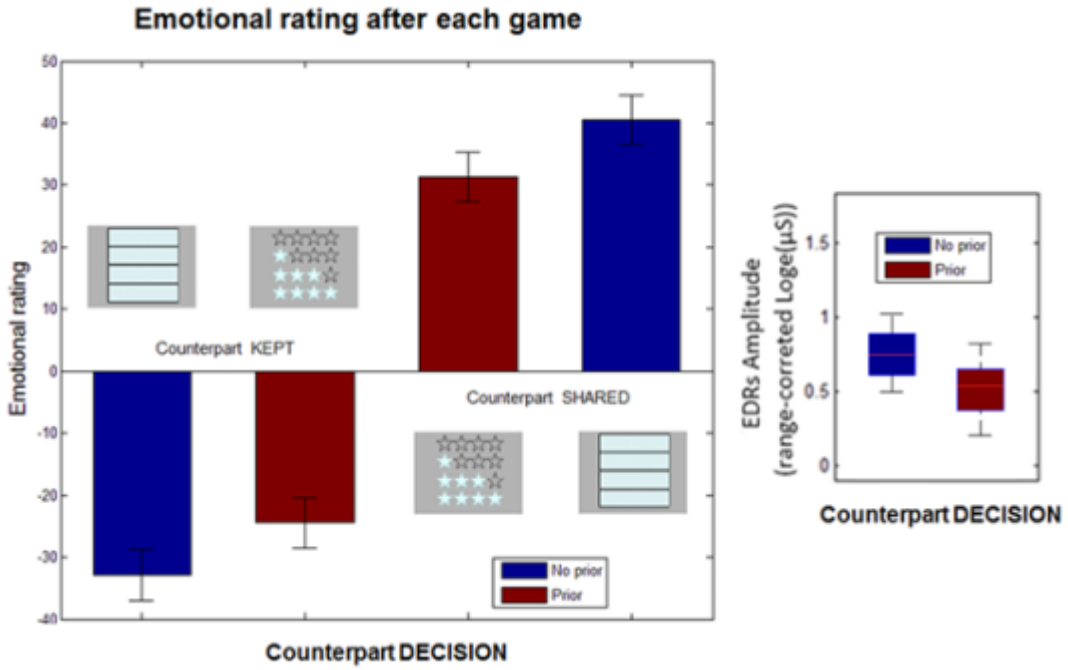


Figure 5.4.3: Differential effect of decision of the participant at time of outcome in the prior and no prior conditions.

habituation across trials (Mann-Whitney test, U = 44856.9, Z = −0.13, p = 0.92). Analysis at time of outcome revealed a main effect of prior on autonomic responses of your participants electrodermal responses. Indeed, EDRs were higher in the no-prior condition (M = 0.742, SE ± 0.098) compare to the prior condition (M = 0.511, SE ± 0.132), (t(41) = 2.01, p=0.01) (**See figure 5.4.3**). We also found that electrodermal responses were correlated with emotional report (r = 0.35, p=0.03). In line with this result and the prior effect, we found that the amplitude of participants' emotional reports were higher in the no prior compare with the prior condition. Indeed, when counterparts shared in the prior condition, participants reported a smaller emotional rate (M = 32.12, SE ± 9.55) than when they shared in the no-prior

condition (M = 41.03, SE ± 8.56), (t=7.0, p<0.001). A similar result was found when counterparts kept (**See figure 5.4.3**). Results also revealed that participants were more affected by inconsistent outcomes compared to consistent outcomes in general. Inconsistent outcomes are computed in two ways: (1) in the prior condition, inconsistencies are calculated as the difference between the expectation of a given decision corresponding to a prior suggestion (i.e. trustworthiness if the counterpart has 3 or 4 stars and untrustworthiness if the counterpart has 0 or 1 stars) and the unexpected outcome (the 3 or 4 stars counterpart violated trust etc.). (2) In the no prior condition, we assumed that players were expecting a trustworthy decision of the counterpart when they decided to share and vice versa when they kept. In that case, the inconsistency was a basic prediction error. Across all inconsistent outcomes, EDR results show higher activity for negative prediction error (M = 0.732, SE ± 0.08) than for positive prediction error (M = 0.387, SE ± 0.04), $t(41) = 3.75$, p=0.001, $\eta^2 = 0.42$. Additionally, when altruistic and cooperative counterparts violated trust, EDRs were higher for the prior condition (3 and 4 stars) (M = 0.876, SE ± 0.07) than for the no-prior condition(M = 0.587, SE = 0.09), $t(41) = 3.05$, p=0.003, $\eta^2 = 0.38$. Within the prior condition, no difference was found between 3 and 4 stars. No difference was also found when counterparts reciprocated trust in the prior and no prior conditions ($t(41) = 0.45$, p=0.67).

5.4.3 Correlation analyses

We found a positive correlation between emotional rating magnitudes (absolute values of emotional reports) and EDRs at time of outcome, $r(41) = 0.52$, p = 0.02. Thus, we decided to further investigate the skin conductance data, integrating it with participant's reported emotions. According to our previous analysis, when investigating inconsistent outcomes, we isolated the outcomes in "Violation-Trust-EDR" and looked at the percentage of correct recognition in Task 2 as well as the rating of trustworthiness. We found a strong correlation between individual magnitude of EDRs after violation of trust and the percentage of trials in which participants remembered the counterparts in second session, $r(41) = 0.63$, p = 0.003 (see figure 5).

Since the correlation between emotional ratings and EDRs yield to a significant correlation, we investigated the difference between emotional rating, trustworthiness rating and recognition in the next task. We found that EDRs after violation of trust were anti-correlated with untrustworthiness ratings ($r(41) = -0.37$, p = 0.053). Finally, we found a positive relationship between the absolute emotional rating in the first task and the recognition in the second task ($r(41) = 0.37$, p = 0.008). This result indicates that the strongest the violation of trust felt by the participants and the more often they were able to remember their counterparts

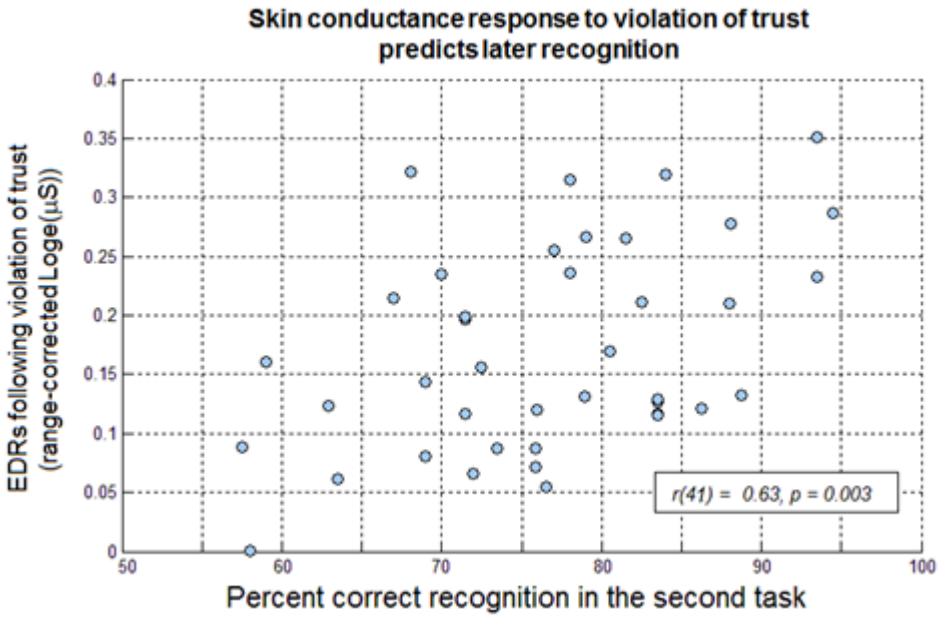


Figure 5.4.4: This plot shows a significant positive correlation between participant's electrodermal responses (EDRs) when seeing that others violated trust and the percentage of correct recognition of these counterparts in the second task (out of a total of 40 trials).

in subsequent tasks. The hierarchical multiple regression analysis indicated that "Violation-Trust-EDR" alone was a significant predictor for later recognition, $F(1,41) = 6.45$, $p = 0.021$; adjusted $R2 = 0.22$ (see table 2). When we added the "Violation-Trust-EDR-Prior" variable in the model, the difference score resulted in an additional 19.78% of variance being explained, $F(2,41) = 7.51$, $p = 0.005$; adjusted $R^2 = 0.42$, $\Delta * R^2 = 0.2$, $p(F\ change) = 0.02$. The "Violation-Trust-EDR-No-Prior" variable didn't contribute to the regression model. The same result was obtained when reversing the entry of the two variables in the model.

5.4.4 Eye tracking preliminary results

Consistent with our prediction, our participants looked at the keys features of the face (facial-AOIs described in method) longer (mean (total fixations) relevant= 1160, s.d. = 190) than at the background region (non-specified regions) (mean (total fixations) no relevant= 199, s.d. = 76) when viewing pictures of their counterparts (t_{30}=30.516, $p < 0.0001$), **See figure 5.4.5**.

Between the prior and the no-prior conditions we found an equal number of saccades (mean (total saccade) no prior = 18132, mean (total saccades) prior = 18009, $t_{30} = 0.61$, $p = 0.54$). However, our results show that participants made more relevant saccades (i.e., a transition between one facial-AOI to another facial-AOI) when encountering a counterparts without

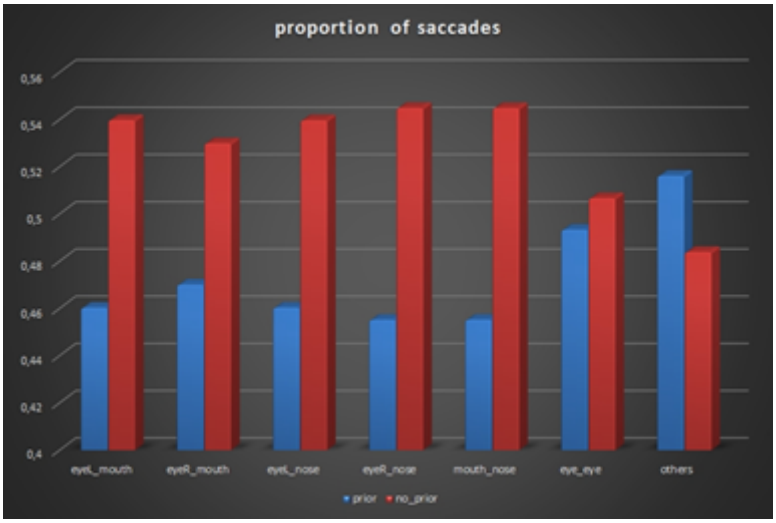


Figure 5.4.5: Participants made more relevant saccades (i.e., a transition between one facial-AOI to another facial-AOI) when encountering a counterparts without priors than when encountering a counterparts with priors.

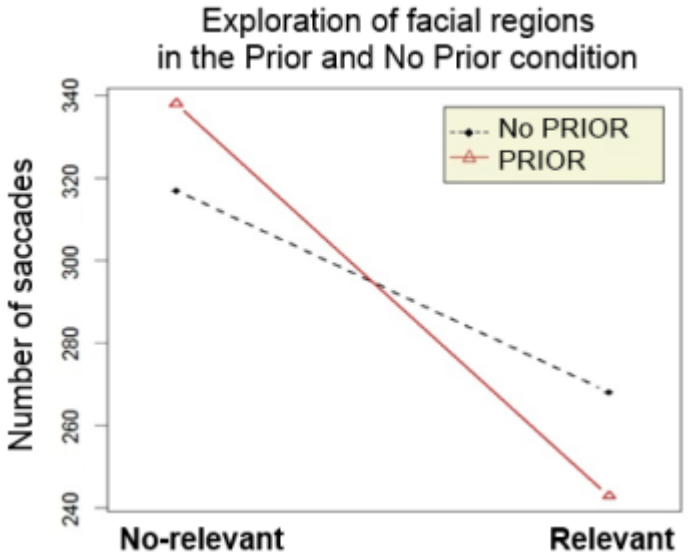


Figure 5.4.6: Cross-interaction between facial-relevance and prior conditions when participants were shown pictures of their counterparts.

priors (Nb = 268, Sd $\pm$ 68) than when encountering a counterparts with priors (Nb = 242, Sd = 75; F_{30} = 15.71, p = 0.0004). This result indicates that participants were exploring more attentively the relevant facials features of others when having no prior information about them. At the same time, it also indicates that the reputational prior diminishes the reliance on facial perception.

To examine the effect of prior and no-prior conditions on relevant and non-relevant saccades, we tested the interactions between saccades relevance and prior's conditions. Non-

relevant saccades were computed as any transition from a facial-AOI to the background - or from the background to a facial-AOI equally -.

A significant cross-interaction was found between facial-relevance and prior's condition when participants were shown pictures of their counterparts (**See figure 5.4.6**). In particular, they made more irrelevant saccades in the prior condition (Nb = 338, Sd $\pm$ 68) compare to the no-prior condition (Nb = 316, Sd $\pm$ 72; F_{30} = 32.79, p = 3^{-06}).

5.5 Conclusions

In this last we recorded electrodermal responses and eye-movements while participants played single-TG with counterparts of various pro-social types. We manipulated the probability of reciprocation of each counterpart that represented their types: competitive (10%), individualistic (30%), cooperative (70%) and altruistic (90%). For half of these counterparts, participants were given reliable reputational priors signaling their types (priors were similar to the ones used in online marketplaces like EBay). Our results show for the first time the effects of reputation the way people evaluate others by observing their face. Indeed, we found that participants paid less attention to the critical parts of the face (eyes, mouth and nose) when they had received a reputation information regarding their counterparts compare to when no prior was provided. This result suggests that reputation allows participants to form a first impression about their counterparts thus paying less attention to their facial-traits. Indeed, when they had received prior information about their counterparts and had to make the decision on whether or not to trust them, we found that the magnitude of their autonomic response was smaller compare to the no-prior condition. This result is in line with others studies showing that physiological emotional responses increase with uncertainty (Epstein and Roupenian, 1970; Critchley, Mathias, and Dolan, 2001). Finally, this study reports for the first time how trust violation affects later recognition. Indeed, our results suggest that the more emotionally aroused participants were when observing untrustworthy outcomes, the more accurate they were at remembering the person that betrayed them in a later task.

We cannot direct the wind, but we can adjust the sails.

Dolly Parton

6

Discussion and conclusion

In this thesis I have presented empirical and theoretical accounts of the decision to trust in social learning and in single-interaction contexts where reputation information is available or not. This work extends previous findings and theories in a number of directions. The work proposes an enhanceda theoretical framework for understanding social decision-making processes and for designing and analyzing future experiments. In particular, richer computational and theoretical models have been introduced which provide new insights and more accurate explanations for behavioral and neural responses in social interactive contexts.

6.1 Summary of contributions

Chapter 3 introduces an Adaptive-Belief RL model that combines learning about relatively long-term social reward expectancies and prior expectations driven by reputation information. A theoretical model has been provided to explain behavioral results of participants playing as trustors in repeated economic games (repeated trust games: RTGs) and to suggest a model of trustworthiness-reputation interactions. The *model-based* formalism used in this new model, along with a non-deterministic update of belief, is clearly more realistic to approach the prob-

ability distribution of reciprocity with parameters that depend both on reputation information and trustworthiness observed during RTGs. To demonstrate the accuracy of the model, we compared its results with those obtained from other RL algorithms that are based on different theoretical principles. In chapter 3, I have also presented other connections between behavioral and computational data suggested by the new model, and found the closest fit to trial-to-trial variability when participants were encountering social inconsistency.

Chapter 4 subsequently expanded the previous study with fMRI techniques, incorporating RTGs with stochastic ending and demonstrates that models of the same family as the one studied in Chapter 3 also fit neural responses to social interactions. We found that the Prior-Expectation model was more relevant for studying RTGs where relationship timing can vary, and this model was used to explain neural patterns recorded when participants made the decision to trust someone and when they updated their beliefs about others. This chapter also exposes a novel hypothesis about the interaction between a prefrontal cortical representational system who's function is to preserve actions suggested by the reputation in case of social inconsistency, and a dopaminergic reward-learning system that continuously tracks reward contingencies. Results support the notion of an "actor/critic" organization in which both social prediction error and action selection are anatomically separated in the brain but functionally correlated (Li, McClure, King-Casas, and Read Montague, 2006; Mahmoudi, Principe, and Sanchez, 2009; Potjans, Morrison, and Diesmann, 2009). FMRI techniques proposed for this interaction combine model-based and model-free reinforcement learning algorithms. The work in this second study also provides important information for understanding how trust and reputation signals are part of a large set of brain systems involved in reducing uncertainty, predicting, learning and maintaining goals.

Chapter 5 embodies a new experimental task designed to probe the consequence of a variety of reputation information and type of partners in behavioral and psychophysiological response changes during TGs as well as in a later recognition task. This new study provides confirmatory and conclusive answers about the effect of reputation in first interactions and of social reward on outcomes evaluations. It presents preliminary data showing that behavioral changes to social inconsistency correlated strongly with sympathetic activity and emotional arousal reports. The most encouraging results explore the relationship between violation of social expectation and memory-recognition and therefore should guide future experimentation and analysis in this area.

6.2 Summary of results

In the context of social decision-making, in particular the decision to trust unfamiliar persons, four main findings are particularly interesting: (i) reputational priors are key social signals to reduce uncertainty at the beginning of a trust-based relationship, (ii) prediction errors underlying social learning mechanisms are modified by reputational priors, (iii) the neural network involved in the decision to trust illustrates the hierarchical role of reputation priors in compensating for social inconsistencies (iv) discrepancies between high-social expectations and trust betrayal influence the recognition-memory system. While (i), (ii) and (iii) allow us to propose new biologically plausible social learning models to illustrate the relationship between reputation and trust, (iv) raises new questions regarding our capacity to trust, learn and remember.

6.2.1 Reputation reduces uncertainty in initial exchanges

Behavioral and psychophysiological evidences

Chapters 3, 4 and 5 present new empirical data confirming that, in the absence of objectively known probabilities of trustworthiness, people exploit reputational cues. Across all three experiments, evidence indicated that people consistently made choices aligned with reputation priors matching the various counterparts' types. In the third study, this effect implied higher psychophysiological (electrodermal) activities in situations of higher uncertainty (when participants had no priors) compared to when reputation information was provided. These results are in line with those reported by other authors on individual and social decision tasks showing that people reduce uncertainty by sampling and exploiting available environmental cues.

fMRI evidences when first a counterpart is presented for the first time

As well as behavioural results, chapter 4 also provides new insights into the role of reputation in modifying brain responses to an unfamiliar person. Enhanced activation in medial-PFC was observed only when reputational priors were provided, regardless of whether they signalled a positive or negative reputation. Along with converging evidence from previous studies, our results validate the role that the medial-PFC plays in guiding social decision-making under uncertainty (Yoshida and Ishii, 2006). In chapter 4, new analyses were used to support this theory showing that the insula displayed higher activity when no reputation information was

available, aligned with previous studies in which insular activity was found to process uncertainty (Preuschoff et al., 2008). Our fMRI investigation supports the notion that these two brain regions, the Medial PFC and the Insula play complementary roles in uncertainty reduction and risk prediction.

Computational evidence of prior-expectation mechanisms

Chapters 3 and 4 propose new RL models of trust-related decisions that reject the assumption from previous models that initial values of social decision-making are allocated randomly (Erev and Roth, 1998). While a large number of previous studies in cognitive science have shown that people use social signals to select initial actions (Li et al., 2011; Delgado et al., 2005; Chang et al., 2010), none had previously incorporated reputation signals in social learning models and acknowledge that these signals have a lot of implications for the performance of the modeled system. The present work bridges the gap by making use of a new version of the RL rule that is equivalent in some aspects to the novelty model proposed by Daw et al in (2010) (Gläscher et al., 2010) but which combines individual sensibilities to uncertainty and the nature of available reputation information. We found that the Prior-Expectation RL model captured most of the inter-individual variability in short relationships and revealed better fits with neural and behavioral activity than the Adaptive-Belief RL model adapted for long-horizon interactions. This result is not surprising considering that our participants had relatively few trials to accumulate sufficient knowledge about the others' types. Thus, when uncertainty about other's type stays rather high, the Prior-Expectation model is more suitable to explain behavioral and neural responses to social interaction valuations.

6.2.2 Reputational priors modify social prediction errors

Behavioral and computational evidences

Another important finding in this work is that computing social reward prediction (RPE) signals similarly than computing primary RPE, such as in animal studies and human probabilistic learning tasks, explain behavioral data better than classical Bayesian learning models (Doll et al., 2012; Beierholm et al., 2011). These results, in conjunction with evidence from human and nonhuman models of reward, suggest that social and primary rewards when inputted into computational models trigger learning in a common fashion. Furthermore, in Chapter 3, I report that the Adaptive-Belief model predicts interaction between social PE and the nature of the reputation. Because the model predicted differential magnitudes for PE when priors were

provided or not, the predictions were more accurate in predicting behaviors than models that do not account for this interaction. While the notion of prior expectation being dynamically shaped by experience is certainly not new, this work proposes for the first time a computational model of this effect in a social context and provides support for its validity. In Chapter 5, I confirm the notion that reputation priors modify outcome-valuation by showing that electrodermal responses to inconsistent behaviors were stronger when following choices aligned with a reputation prior compared to unbiased choices. Indeed, when counterparts with altruistic or cooperative reputations violated trust, electrodermal activity was higher than when counterparts had no reputations.

Neural response to social prediction error with and without priors

In Chapter 4 I present differential activities in the caudate nucleus (part of the reward circuitry) in mediating trial-to-trial adjustments to PE when reputational priors were provided or not. Indeed, using parametric analysis of trial-to-trial estimated PE, we showed that bilateral caudate nuclei were encoding RTG outcomes as an unbiased reward prediction problem only when no priors were given. In fact, in the prior condition, dorso-striatal activity minimally tracked computed-PE and responded selectively to violation of trust. Although this brain area was not found to capture the trial-to-trial variability suggested by the RL model, activation in the dorso-striatum was maximal when a counterpart with a cooperative reputation violated trust. In chapter 4, I argue that the dorso-striatum is still capturing a conflicting signal between expectation and reality and such conflict is actually maximized in the scenario of violation of trust by a partner with a high pro-social reputation. Indeed, when a particular prior is provided, the expectation of a corresponding behavioral is stronger and so should be the discrepancy between this expectation and the negative outcomes. Taken together, the results reported in Chapter 4 suggest that although reputational priors moderate trial-to-trial variability of reward-based learning mechanisms in the brain, the neural signatures of conflict monitoring remain unchanged and even enhanced when reputation priors are present.

Experience attenuate reputation priors

In Chapters 3 and 4, I report that accumulated observations diminish initial reputation influences over time across all RL modeled data fitting behavioral responses. These results support and extend previous findings which have found shown that, although priors modify PE in uncertain environment, experience accumulated over time without initial bias overwhelm beliefs suggested by social cues in risky choices (Cohen, 2006; Huettel et al., 2005). As we have seen

in chapter 3 and 4, the properties of social feedback-based valuation emerged robustly over time even if reputational priors influenced early valuation and have a sustainable effect on outcome-decision. These findings underline the well-established role of the reward circuitry (and in particular the role of the striatum) in encoding decision-outcome associations for target rewards (Kable and Glimcher, 2007; Fareri et al., 2012; Schönberg et al., 2007) and enlarge its role in triggering learning to the social domain. The results of this work fall very much in line with other recent studies (Gläscher et al., 2010; Daw, Gershman, Seymour, Dayan, and Dolan, 2011)

6.2.3 Trust and reputation dynamic

Behavioral and psychophysiological evidences

In chapters 3 and 4 I report a robust and consistent interaction between valuation of outcomes and subsequent decisions aligned with or without reputation information. More specifically, we found that trust-violations damaged interpersonal trust more when the decision to trust was made without reputational bias compared to when the decision was aligned with reputation. This intriguing result is counter-intuitive given the results that I have just presented. Indeed, while we found greater emotional and neural responses to negative inconsistency in the prior condition, participants neglected this conflict-related signal and were persistent with their initial decision in subsequent trials. In fact, general evidences from behavioral economic and cognitive decision-making research have established that the outcome of one decision influences the subsequent decision (McCabe et al., 2001; King-Casas et al., 2005; Krueger et al., 2007), but this study is the first to show that initial reputation-related information mediate these adjustments in behavioral strategies. I argue that (i) decision-making is a dynamic process interfacing several sub-processes (evaluating the outcome, adjusting future behavior, pursuing goals etc.) that can conflict with one another and (ii) that brain patterns involved in such sub-processes would exhibit activity that reflects strategic adjustments in behavior occurring as a result of the evaluation of conflicting signals. In the next paragraph, I review our fMRI results on the topic.

vLPFC modulates the reward-related signal in the dorso-striatum

Chapter 4 presents one of the most unexpected and intriguing result found during our fMRI investigation. We have seen that the striatum was strongly deactivated when violation of trust happened, consistent with its well-described spontaneous changes to unexpected outcomes.

However, we discovered that a high functional connectivity between the caudate nucleus and the lateral prefrontal cortex prevented participants from retaliation after violation of trust in the prior condition. In fact, the strength of connectivity between the dorsal striatum and vLPFC was greater when priors were provided than when they were not. This result suggests a role of the prefrontal cortico-striatal loop in modulating goal directed behavior (in our study, reputational prior were reliable predictors to optimum strategies) in line with previous studies (Valentin et al., 2007; Souza et al., 2009). We thus propose that the vLPFC contributes in maintaining choices aligned with the reputational prior when expectations conflict with reality. This finding supports recent empirical evidences showing that the vLPFC exerts a top-down cognitive modulation of other brain patterns toward contextually appropriate beliefs (Cohen et al., 1990; Miller and Cohen, 2001). We argue that the vLPFC, a phylogenetically younger brain regions involved in higher cognition, may contribute to maintain beliefs and decisions anchored to reputational priors, thus relatively discount the weight of conflicting evidence. The interplay between striatum and vLPFC may prevent unnecessary retaliation when others violate our trust, and thus constitutes an important neurocognitive modulator that favors social stability.

6.2.4 Violation of trust influences the recognition-memory system

In the last study presented in this thesis, we investigated the link between a person's autonomic responses (EDRs) when observing trust violation and later recognition of the person responsible for that damage. As a preliminary result we observed, for the first time, that inter-individual differences in spontaneous EDRs predicted later recognition of counterparts that violated trust. This finding relates to previous studies which probed inter-individual galvanic responses variability, emotional ratings and social evaluation (Critchley, 2002). Specifically, we found that the magnitude of EDRs when observing trust violation was correlated with people's differential emotional ratings, confirming that sympathetic activity reflects the vicarious emotional reports to social outcomes. Taken together, these encouraging preliminary results support the assumption that damage to trust has consequences in the reward and memory system.

Finally, the work presented here may guides future investigations onf a number of issues, some of which I now review.

6.3 Future directions

Many questions are raised by the computational theories and results proposed here that could fruitfully be investigated in future theoretical and experimental studies.

1. The experimental task proposed in this work is based on a simplified version of the trust game that allows a clear distinction between choices (trust/no trust) and similarly, a clear distinction between opponent's responses (trustworthy/untrustworthy). However, it would be interesting to use the original version of the trust game, in future studies manipulating priors, in which participants can invest any amount of their initial endowment (for example from 0 to 20 euros by increment of 1) to the other player. Such experimental set-ups would allow for a rather precise investigation of how starting values of RL models impact the estimated parameters of the update rule. Such a design would also specify the behavior of the Adaptive-Prior and the Prior-Expectation models. Analyzing how the trial-to-trial variation of money invested in each RTG would allow confirming or rejecting the proposed models in social contexts.

2. In addition, hierarchical reinforcement learning model in which the suggestion carried by reputation is transferred to a sub-process that trigger learning from direct feedback could be rather tested instead of pure Markovian (or semi-Markovian) RL processes sampling events through beliefs and time dimensions. Hierarchical RL models treat problem solving by introducing representation (or beliefs) into the planning/expecting systems. These types of models allow the notion of a hierarchy between beliefs and updates functions and the possibility for a certain representation to evoke within a single action-goal, an entire sequences of possible actions.

3. In (2001) Knutson et al. underlined the important aspect of representation of costs when decisions are made that lead to a broad new series of studies exploring cost-value decisions (Knutson et al., 2001). In the domain of social decision-making however, none have explored the cost to obtain reputation signals in order to make rapid and costless decision in first interactions. For example, in an experimental setting closed to EBay's type of website where reputation is a key feature to consider before engaging in a transaction, future studies could explore the amount of money people would be willing to pay in order to access reputational information. Not only this type of studies would provide new insights on how people explore and exploit reputation-cues in order to engage in economical transactions but it will also help us understand how the brain integrates costs and valuable-cues during social decision-making.

4. In addition, the reliability of the prior seems to be an extra sensible parameter to consider. In fact, previous studies using social information as proxy for reputation signals (information about moral character of others, cultural attributes, advices, instructions etc.) also found an effect of those signals on decisions, even when they did not convey any reliable information (e.g. the simulation of partner's responses were randomly distributed). Our results, complementing these previous studies, suggest that the reliability of pro-social signals could be investigated as an independent variable in studies and models that probe adaptive social decision-making. For example, still in an EBay type of scenario, the reliability of the reputation would be the amount of persons that gave feedback scores for a certain seller. If just one person gave report about a seller's trustworthiness the rating is less valuable and reliable than if 100 persons contribute to the rating. Future studies would benefit in taking into consideration these types of uncertainty, including uncertainty about the providers of reputation information.

6.4 General conclusion

In this thesis, I provide a general background and formalism for studying trust and reputation in humans using a combination of techniques such as game theoretical experiments, computational models, functional neuroimaging and electrodermal recording. Describing social interaction with model-free and model-based reinforcement learning processes, I present new models allowing robustness and efficient patterns in risky environments that fit behavioral and neural responses of participants engaged in TGs. This work delves beyond previous research in its consideration of the complexities of assessing environmental cues, weighting options, making decisions and evaluating social outcomes with respect to trust and reputation.

References

R Adolphs, D Tranel, and A R Damasio. The human amygdala in social judgment. *Nature*, 393 (6684):470–474, June 1998. ISSN 0028-0836. doi: 10.1038/30982. PMID: 9624002.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, December 1974. ISSN 0018-9286. doi: 10.1109/TAC.1974.1100705.

Jose Apesteguia, Steffen Huck, and Jörg Oechssler. Imitation—theory and experimental evidence. *Journal of Economic Theory*, 136(1):217–235, September 2007. ISSN 0022-0531. doi: 10.1016/j.jet.2006.07.006.

Kenneth J. Arrow. Gifts and exchanges. *Philosophy and Public Affairs*, 1(4):343–362, 1972.

R Axelrod and W D Hamilton. The evolution of cooperation. *Science (New York, N.Y.)*, 211 (4489):1390–1396, March 1981. ISSN 0036-8075. PMID: 7466396.

Bernard W Balleine. Neural bases of food-seeking: affect, arousal and reward in corticostriatolimbic circuits. *Physiology & behavior*, 86(5):717–730, December 2005. ISSN 0031-9384. doi: 10.1016/j.physbeh.2005.08.061. PMID: 16257019.

Bernard W. Balleine. The neural basis of choice and decision making. *The Journal of Neuroscience*, 27(31):8159–8160, August 2007. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.1939-07.2007.

Moshe Bar, Maital Neta, and Heather Linz. Very first impressions. *Emotion (Washington, D.C.)*, 6(2):269–278, May 2006. ISSN 1528-3542. doi: 10.1037/1528-3542.6.2.269. PMID: 16768559.

Dominic J. Barraclough, Michelle L. Conroy, and Daeyeol Lee. Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience*, 7(4):404–410, March 2004. doi: 10.1038/nn1209.

Greg Barron and Ido Erev. Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, 16(3):215–233, 2003. ISSN 1099-0771. doi: 10.1002/bdm.443.

Nadège Bault, Giorgio Coricelli, and Aldo Rustichini. Interdependent utilities: How social ranking affects choice behavior. *PLoS ONE*, 3(10):e3477, October 2008. doi: 10.1371/journal.pone.0003477.

Thomas Baumgartner, Markus Heinrichs, Aline Vonlanthen, Urs Fischbacher, and Ernst Fehr. Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron*, 58 (4):639–650, May 2008. ISSN 1097-4199. doi: 10.1016/j.neuron.2008.04.009. PMID: 18498743.

Hannah M. Bayer and Paul W. Glimcher. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, 47(1):129–141, July 2005. ISSN 0896-6273. doi: 10.1016/j.neuron.2005.05.020. PMID: 15996553 PMCID: PMC1564381.

Timothy E. J. Behrens, Laurence T. Hunt, Mark W. Woolrich, and Matthew F. S. Rushworth. Associative learning of social value. *Nature*, 456(7219):245–249, November 2008. ISSN 0028-0836. doi: 10.1038/nature07538.

Ulrik R. Beierholm, Cedric Anen, Steven Quartz, and Peter Bossaerts. Separate encoding of model-based and model-free valuations in the human brain. *NeuroImage*, 58(3):955–962, October 2011. ISSN 10538119. doi: 10.1016/j.neuroimage.2011.06.071.

Goedele van Belle, Meike Ramon, Philippe Lefèvre, and Bruno Rossion. Fixation patterns during recognition of personally familiar and unfamiliar faces. *Frontiers in Cognitive Science*, 1:20, 2010. doi: 10.3389/fpsyg.2010.00020.

Richard Ernest Bellman. *Dynamic Programming*. Dover, 2003. ISBN 9780486428093.

J. Berg, J. Dickhaut, and K. McCabe. Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122–142, 1995.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 4th edition, June 2012. ISBN 1886529086.

Guido Biele, Jörg Rieskamp, and Richard Gonzalez. Computational models for the combination of advice and individual learning. *Cognitive Science*, 33(2):206–242, March 2009. ISSN 03640213, 15516709. doi: 10.1111/j.1551-6709.2009.01010.x.

Guido Biele, Jörg Rieskamp, Lea K. Krugel, and Hauke R. Heekeren. The neural basis of following advice. *PLoS Biol*, 9(6):e1001089, June 2011. doi: 10.1371/journal.pbio.1001089.

Munindar P. Singh Bin Yu. Detecting deception in reputation management. 2002.

Riccardo Boero, Giangiacomo Bravo, Marco Castellani, and Flaminio Squazzoni. Reputational cues in repeated trust games. *The Journal of Socio-Economics*, 38(6):871–877, December 2009. ISSN 1053-5357. doi: 10.1016/j.socec.2009.05.004.

Iris Bohnet and Yael Baytelman. Institutions and trust implications for preferences, beliefs and behavior. *Rationality and Society*, 19(1):99–135, February 2007. ISSN 1043-4631, 1461-7358. doi: 10.1177/1043463107075110.

Iris Bohnet and Steffen Huck. Repetition and reputation: Implications for trust and trustworthiness when institutions change. *The American Economic Review*, 94(2):362–366, May 2004. ISSN 0002-8282. doi: 10.2307/3592911. ArticleType: research-article / Issue Title: Papers and Proceedings of the One Hundred Sixteenth Annual Meeting of the American Economic Association San Diego, CA, January 3-5, 2004 / Full publication date: May, 2004 / Copyright © 2004 American Economic Association.

Iris Bohnet and Richard Zeckhauser. Trust, risk and betrayal. Working Paper Series rwp03-041, Harvard University, John F. Kennedy School of Government, 2003.

S Boon and J Holmes. The dynamics of interpersonal trust: Resolvinguncertainty in the face of risk. *Cooperation and prosocial behaviour*, pages 190–211, 1991.

Wouter van den Bos, Eric van Dijk, and Eveline A. Crone. Learning whom to trust in repeated social interactions: A developmental perspective. *Group Processes & Intergroup Relations*, 15(2):243–256, March 2012. ISSN 1368-4302, 1461-7188.

Robert Boyd. Mistakes allow evolutionary stability in the repeated prisoner's dilemma game. *Journal of Theoretical Biology*, 136(1):47–56, January 1989. ISSN 0022-5193. doi: 10.1016/S0022-5193(89)80188-2.

Ethan S Bromberg-Martin, Masayuki Matsumoto, Simon Hong, and Okihide Hikosaka. A pallidus-habenula-dopamine pathway signals inferred stimulus values. *Journal of neurophysiology*, 104(2):1068–1076, August 2010. ISSN 1522-1598. doi: 10.1152/jn.00158.2010. PMID: 20538770.

D. B. Bromley. *Reputation, Image and Impression Management*. John Wiley & Sons, June 1993. ISBN 9780471938699.

D B Bromley. Psychological aspects of corporate identity, image and reputation. *Corporate Reputation Review*, 3(3):240–252, July 2000. ISSN 1363-3589, 1479-1889. doi: 10.1057/palgrave.crr.1540117.

Silvia A Bunge. How we use rules to select actions: a review of evidence from cognitive neuroscience. *Cognitive, affective & behavioral neuroscience*, 4(4):564–579, December 2004. ISSN 1530-7026. PMID: 15849898.

George Bush, Phan Luu, and Michael I. Posner. Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences*, 4(6):215–222, June 2000. ISSN 1364-6613. doi: 10.1016/S1364-6613(00)01483-2.

Luis M. B. Cabral. The economics of trust and reputation: A primer. In press.

Colin Camerer and Keith Weigelt. Experimental tests of a sequential equilibrium reputation model. *Econometrica*, 56(1):1–36, January 1988. ISSN 0012-9682. doi: 10.2307/1911840. ArticleType: research-article / Full publication date: Jan., 1988 / Copyright © 1988 The Econometric Society.

Colin F. Camerer. Neuroeconomics: Using neuroscience to make economic predictions*. *The Economic Journal*, 117(519):C26–C42, 2007. ISSN 1468-0297. doi: 10.1111/j.1468-0297.2007.02033.x.

Samuele Centorrino, Elodie Djemai, Astrid Hopfensitz, Manfred Milinski, and Paul Seabright. Smiling is a costly signal of cooperation opportunities: Experimental evidence from a trust game. SSRN Scholarly Paper ID 1846256, Social Science Research Network, Rochester, NY, May 2011.

Luke J. Chang and Alan G. Sanfey. Unforgettable ultimatums? expectation violations promote enhanced social memory following economic bargaining. *Frontiers in Behavioral Neuroscience*, 3, October 2009. ISSN 1662-5153. doi: 10.3389/neuro.08.036.2009. PMID: 19876405 PMCID: PMC2769546.

Luke J. Chang, Bradley B. Doll, Mascha van 't Wout, Michael J. Frank, and Alan G. Sanfey. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology*, 61(2):87–105, September 2010. ISSN 00100285. doi: 10.1016/j.cogpsych.2010.03.001.

Jonathan D. Cohen, James L. Mcclelland, and Kevin Dunbar. On the control of automatic processes: A parallel distributed processing account of the stroop effect. *Psychological Review*, 97:332–361, 1990.

M X Cohen, A S Heller, and C Ranganath. Functional connectivity with anterior cingulate and orbitofrontal cortices during decision-making. *Brain research. Cognitive brain research*, 23(1):61–70, April 2005. ISSN 0926-6410. doi: 10.1016/j.cogbrainres.2005.01.010. PMID: 15795134.

Michael X. Cohen. Individual differences and the neural representations of reward expectation and reward prediction error. *Social Cognitive and Affective Neuroscience*, 2(1):20–30, March 2006. ISSN 1749-5016, 1749-5024. doi: 10.1093/scan/nsl021.

Michael X. Cohen and Charan Ranganath. Reinforcement learning signals predict future decisions. *The Journal of Neuroscience*, 27(2):371–378, January 2007. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.4421-06.2007.

Coleman. Foundations of social theory — james coleman | harvard university press, 1994.

Thomas F. Coleman and Yuying Li. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on Optimization*, 6(2):418–445, May 1996. ISSN 1052-6234, 1095-7189. doi: 10.1137/0806023.

James C. Cox. How to identify trust and reciprocity. *Games and Economic Behavior*, 46(2): 260–281, February 2004. ISSN 0899-8256. doi: 10.1016/S0899-8256(03)00119-2.

Hugo D. Critchley. Book review: Electrodermal responses: What happens in the brain. *The Neuroscientist*, 8(2):132–142, April 2002. ISSN 1073-8584, 1089-4098.

Hugo D Critchley, Christopher J Mathias, and Raymond J Dolan. Neural activity in the human brain relating to uncertainty and arousal during anticipation. *Neuron*, 29(2):537–545, February 2001. ISSN 0896-6273.

Kim M. Dalton, Brendon M. Nacewicz, Tom Johnstone, Hillary S. Schaefer, Morton Ann Gernsbacher, H. H. Goldsmith, Andrew L. Alexander, and Richard J. Davidson. Gaze fixation and the neural circuitry of face processing in autism. *Nature Neuroscience*, 8(4): 519–526, 2005. ISSN 1097-6256. doi: 10.1038/nn1421.

Nathaniel D Daw and Kenji Doya. The computational neurobiology of learning and reward. *Current opinion in neurobiology*, 16(2):199–204, April 2006. ISSN 0959-4388. doi: 10.1016/j.conb.2006.03.006. PMID: 16563737.

Nathaniel D. Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12): 1704–1711, December 2005. ISSN 1097-6256. doi: 10.1038/nn1560.

Nathaniel D. Daw, Samuel J. Gershman, Ben Seymour, Peter Dayan, and Raymond J. Dolan. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6): 1204–1215, March 2011. ISSN 0896-6273. doi: 10.1016/j.neuron.2011.02.027.

Peter Dayan and Yael Niv. Reinforcement learning: the good, the bad and the ugly. *Current opinion in neurobiology*, 18(2):185–196, April 2008. ISSN 0959-4388. doi: 10.1016/j.conb.2008.08.003. PMID: 18708140.

Dominique J-F de Quervain, Urs Fischbacher, Valerie Treyer, Melanie Schellhammer, Ulrich Schnyder, Alfred Buck, and Ernst Fehr. The neural basis of altruistic punishment. *Science (New York, N.Y.)*, 305(5688):1254–1258, August 2004. ISSN 1095-9203. doi: 10.1126/science.1100735. PMID: 15333831.

Jean Decety, Philip L Jackson, Jessica A Sommerville, Thierry Chaminade, and Andrew N Meltzoff. The neural bases of cooperation and competition: an fMRI investigation. *NeuroImage*, 23(2):744–751, October 2004. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2004.05.025. PMID: 15488424.

Thomas Deelmann and Peter Loos. Trust economy: Aspects of reputation and trust building for smes in e-business. *AMCIS 2002 Proceedings*, December 2002.

M R Delgado, R H Frank, and E A Phelps. Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature neuroscience*, 8(11):1611–1618, November 2005. ISSN 1097-6256. doi: 10.1038/nn1575. PMID: 16222226.

A. Dickinson. Actions and habits: The development of behavioural autonomy. *Royal Society of London Philosophical Transactions Series B*, 308:67–78, February 1985. ISSN 0962-8436. doi: 10.1098/rstb.1985.0010.

Anthony Dickinson and Bernard Balleine. The role of learning in the operation of motivational systems. In *Stevens' Handbook of Experimental Psychology*. John Wiley & Sons, Inc., 2002. ISBN 9780471214427.

Bradley B Doll, W Jake Jacobs, Alan G Sanfey, and Michael J Frank. Instructional control of reinforcement learning: a behavioral and neurocomputational investigation. *Brain research*, 1299:74–94, November 2009. ISSN 1872-6240. doi: 10.1016/j.brainres.2009.07.007. PMID: 19595993.

Bradley B. Doll, Kent E. Hutchison, and Michael J. Frank. Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *The Journal of Neuroscience*, 31(16):6188–6198, April 2011. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.6486-10.2011.

Bradley B Doll, Dylan A Simon, and Nathaniel D Daw. The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, 22(6):1075–1081, December 2012. ISSN 0959-4388. doi: 10.1016/j.conb.2012.08.003.

Dominic S Fareri, Luke J. Chang, and Mauricio R. Delgado. Effects of direct social experience on trust decisions and neural. *Front Neurosci.*, 6: 148, 2012. doi: 10.3389/fnins.2012.00148.

Kenji Doya. Modulators of decision making. *Nature neuroscience*, 11(4):410–416, April 2008. ISSN 1097-6256. doi: 10.1038/nn2077. PMID: 18368048.

Barnaby D Dunn, Davy Evans, Dasha Makarova, Josh White, and Luke Clark. Gut feelings and the reaction to perceived inequity: the interplay between bodily responses, regulation, and perception shapes the rejection of unfair offers on the ultimatum game. *Cognitive, affective & behavioral neuroscience*, 12(3):419–429, September 2012. ISSN 1531-135X. doi: 10.3758/s13415-012-0092-z. PMID: 22618636.

Catherine Eckel and Rick Wilson. Is trust a risky decision? SSRN Scholarly Paper ID 1843473, Social Science Research Network, Rochester, NY, May 2011.

D M Egelman, C Person, and P R Montague. A computational role for dopamine delivery in human decision-making. *Journal of cognitive neuroscience*, 10(5):623–630, September 1998. ISSN 0898-929X. PMID: 9802995.

Hedwig Eisenbarth and Georg W Alpers. Happy mouth and sad eyes: scanning emotional facial expressions. *Emotion (Washington, D.C.)*, 11(4):860–865, August 2011. ISSN 1931-1516. doi: 10.1037/a0022758. PMID: 21859204.

Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, May 1992. doi: 10.1080/02699939208411068.

Andrew D. Engell, James V. Haxby, and Alexander Todorov. Implicit trustworthiness decisions: Automatic coding of face properties in the human amygdala. *Journal of Cognitive Neuroscience*, 19(9):1508–1519, August 2007. ISSN 0898-929X. doi: 10.1162/jocn.2007.19.9.1508.

Seymour Epstein and Armen Roupenian. Heart rate and skin conductance during experimentally induced anxiety: The effect of uncertainty about receiving a noxious stimulus. *Journal of Personality and Social Psychology*, 16:20–28, 1970. ISSN 00223514.

Ido Erev and Alvin E. Roth. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review*, 88 (4):848–81, 1998.

Monique Ernst and Martin P. Paulus. Neurobiology of decision making: A selective review from a neurocognitive and clinical perspective. *Biological Psychiatry*, 58(8):597–604, October 2005. ISSN 0006-3223. doi: 10.1016/j.biopsych.2005.06.004.

Karleyton C Evans, Christopher I Wright, Michelle M Wedig, Andrea L Gold, Mark H Pollack, and Scott L Rauch. A functional MRI study of amygdala responses to angry schematic faces in social anxiety disorder. *Depression and anxiety*, 25(6):496–505, 2008. ISSN 1520-6394. doi: 10.1002/da.20347. PMID: 17595018.

Armin Falk and Urs Fischbacher. A theory of reciprocity. *Games and Economic Behavior*, 54 (2):293–315, February 2006. ISSN 0899-8256. doi: 10.1016/j.geb.2005.03.001.

Dominic S. Fareri, Luke J. Chang, and Mauricio R. Delgado. Effects of direct social experience on trust decisions and neural reward circuitry. *Frontiers in Neuroscience*, 6, October 2012. ISSN 1662-4548. doi: 10.3389/fnins.2012.00148. PMID: 23087604 PMCID: PMC3472892.

Ernst Fehr and Colin F. Camerer. Social neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Sciences*, 11(10):419–427, October 2007. ISSN 1364-6613. doi: 10.1016/j.tics.2007.09.002.

Ernst Fehr, Urs Fischbacher, and Michael Kosfeld. Neuroeconomic foundation of trust and social preferences. SSRN Scholarly Paper ID 781430, Social Science Research Network, Rochester, NY, August 2005.

Elsa Fouragnan, Gabriele Chierchia, Susanne Greiner, Remi Neveu, Paolo Avesani, and Giorgio Coricelli. Reputational priors magnify striatal responses to violations of trust. *The Journal of Neuroscience*, 33(8):3602–3611, February 2013. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.3086-12.2013.

Michael J Frank, Randall C O'Reilly, and Tim Curran. When memory fails, intuition reigns: midazolam enhances implicit inference in humans. *Psychological science*, 17(8):700–707, August 2006. ISSN 0956-7976. doi: 10.1111/j.1467-9280.2006.01769.x. PMID: 16913953.

Wendy L. Fraser. *Trust Violation and Repair: An Exploration of the Views of Work Group Members.* Proquest, Umi Dissertation Publishing, September 2011. ISBN 1243721049.

L W Freedman, A S Scerbo, M E Dawson, A Raine, W O McClure, and P H Venables. The relationship of sweat gland count to electrodermal activity. *Psychophysiology*, 31(2):196–200, March 1994. ISSN 0048-5772. PMID: 8153256.

James Friedrich. Primary error detection and minimization (PEDMIN) strategies in social cognition: A reinterpretation of confirmation bias phenomena. *Psychological Review*, 100 (2):298–319, 1993. ISSN 0033-295X. doi: 10.1037/0033-295X.100.2.298.

K J Friston, C Buechel, G R Fink, J Morris, E Rolls, and R J Dolan. Psychophysiological and modulatory interactions in neuroimaging. *NeuroImage*, 6(3):218–229, October 1997. ISSN 1053-8119. doi: 10.1006/nimg.1997.0291. PMID: 9344826.

Drew Fudenberg, David M. Kreps, and Eric S. Maskin. Repeated games with long-run and short-run players. *The Review of Economic Studies*, 57(4):555–573, October 1990. ISSN 0034-6527. doi: 10.2307/2298086. ArticleType: research-article / Full publication date: Oct., 1990 / Copyright © 1990 The Review of Economic Studies, Ltd.

Paul W. Glimcher, Ernst Fehr, Colin Camerer, and Russell Alan Poldrack, editors. *Neuroeconomics: Decision Making and the Brain*. Academic Press, 1 edition, October 2008. ISBN 0123741769.

Jan Gläscher, Nathaniel Daw, Peter Dayan, and John P. O'Doherty. States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4):585–595, May 2010. ISSN 0896-6273. doi: 10.1016/j.neuron.2010.04.016.

Piotr J. Gmytrasiewicz and Prashant Doshi. A framework for sequential planning in multi-agent settings. *J. Artif. Int. Res.*, 24(1):49–79, July 2005. ISSN 1076-9757.

Cleotilde Gonzalez, Jason Dana, Hideya Koshino, and Marcel Just. The framing effect and risky decisions: Examining cognitive functions with fMRI. *Journal of Economic Psychology*, 26(1):1–20, 2005.

Adam J. Guastella, Philip B. Mitchell, and Mark R. Dadds. Oxytocin increases gaze to the eye region of human faces. *Biological Psychiatry*, 63(1):3–5, January 2008. ISSN 0006-3223.

Alan N. Hampton, Peter Bossaerts, and John P. O'Doherty. Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences*, April 2008. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0711099105.

Ralph Hertwig, Greg Barron, Elke U Weber, and Ido Erev. Decisions from experience and the effect of rare events in risky choice. *Psychological science*, 15(8):534–539, August 2004. ISSN 0956-7976. doi: 10.1111/j.0956-7976.2004.00715.x. PMID: 15270998.

Daeyeol Lee Hiroshi Abe. Distributed coding of actual and hypothetical outcomes in the orbital and dorsolateral prefrontal cortex. *Neuron*, 70(4):731–41, 2011. ISSN 1097-4199. doi: 10.1016/j.neuron.2011.03.026.

Arne Öhman. Automaticity and the amygdala: Nonconscious responses to emotional faces. *Current Directions in Psychological Science*, 11(2):62–66, April 2002. ISSN 0963-7214, 1467-8721. doi: 10.1111/1467-8721.00169.

Clay B Holroyd and Michael G H Coles. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological review*, 109 (4):679–709, October 2002. ISSN 0033-295X. PMID: 12374324.

Kesseley Hong and Iris Bohnet. Status and distrust: The relevance of inequality and betrayal aversion. Working Paper Series rwp04-041, Harvard University, John F. Kennedy School of Government, 2004.

Daniel Houser, Daniel Schunk, and Joachim Winter. Distinguishing trust from risk: An anatomy of the investment game. *Journal of Economic Behavior & Organization*, 74(1–2): 72–81, May 2010. ISSN 0167-2681. doi: 10.1016/j.jebo.2010.01.002.

Ming Hsu, Meghana Bhatt, Ralph Adolphs, Daniel Tranel, and Colin F. Camerer. Neural systems responding to degrees of uncertainty in human decision-making. *Science*, 310(5754): 1680–1683, December 2005. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1115327.

Scott A. Huettel, Allen W. Song, and Gregory McCarthy. Decisions under uncertainty: Probabilistic context influences activation of prefrontal and parietal cortices. *The Journal of Neuroscience*, 25(13):3304–3311, March 2005. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.5070-04.2005.

R. Mark Isaac, Kenneth F. McCue, and Charles R. Plott. Public goods provision in an experimental environment. *Journal of Public Economics*, 26(1):51–74, 1985. ISSN 0047-2727.

Julian Jamison, Dean S. Karlan, and Laura Schechter. To deceive or not to deceive: The effect of deception on behavior in future laboratory experiments. SSRN Scholarly Paper ID 913057, Social Science Research Network, Rochester, NY, June 2006.

Ryan K Jessup, Anthony J Bishara, and Jerome R Busemeyer. Feedback produces divergence from prospect theory in descriptive choice. *Psychological science*, 19(10):1015–1022, October 2008. ISSN 1467-9280. doi: 10.1111/j.1467-9280.2008.02193.x. PMID: 19000212.

Adam Johnson, Matthijs AA van der Meer, and A David Redish. Integrating hippocampus and striatum in decision-making. *Current Opinion in Neurobiology*, 17(6):692–697, December 2007. ISSN 09594388. doi: 10.1016/j.conb.2008.01.003. URL http://linkinghub.elsevier.com/retrieve/pii/S0959438808000056.

Noel Johnson and Alexandra Mislin. Trust games: A meta-analysis. SSRN Scholarly Paper ID 1702678, Social Science Research Network, Rochester, NY, November 2010.

Rebecca M. Jones, Leah H. Somerville, Jian Li, Erika J. Ruberry, Victoria Libby, Gary Glover, Henning U. Voss, Douglas J. Ballon, and BJ Casey. Behavioral and neural properties of social reinforcement learning. *The Journal of Neuroscience*, 31(37):13039–13045, September 2011. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.2972-11.2011. PMID: 21917787 PMCID: PMC3303166.

Joseph W Kable and Paul W Glimcher. The neural correlates of subjective value during intertemporal choice. *Nature neuroscience*, 10(12):1625–1633, December 2007. ISSN 1097-6256. doi: 10.1038/nn2007. PMID: 17982449.

Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore. Reinforcement learning: a survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

Sham Kakade and Peter Dayan. Dopamine: generalization and bonuses. *Neural networks: the official journal of the International Neural Network Society*, 15(4-6):549–559, July 2002. ISSN 0893-6080. PMID: 12371511.

Yasumasa Ueda Kazuyuki Samejima. Representation of action-specific reward values in the striatum. *Science (New York, N.Y.)*, 310(5752):1337–40, 2005. ISSN 1095-9203. doi: 10.1126/science.1115270.

Harris H. Kim. *Market uncertainty and socially embedded reputation.(Report): An article from: The American Journal of Economics and Sociology*. Blackwell Publishers Ltd., July 2009.

Brooks King-Casas, Damon Tomlin, Cedric Anen, Colin F Camerer, Steven R Quartz, and P Read Montague. Getting to know you: reputation and trust in a two-person economic exchange. *Science (New York, N.Y.)*, 308(5718):78–83, April 2005. ISSN 1095-9203. doi: 10.1126/science.1108062. PMID: 15802598.

B Knutson, C M Adams, G W Fong, and D Hommer. Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 21(16):RC159, August 2001. ISSN 1529-2401. PMID: 11459880.

Amy L Krain, Amanda M Wilson, Robert Arbuckle, F Xavier Castellanos, and Michael P Milham. Distinct neural mechanisms of risk and ambiguity: a meta-analysis of decision-making. *NeuroImage*, 32(1):477–484, August 2006. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2006.02.047. PMID: 16632383.

Konrad Körding. Decision theory: what "should" the nervous system do? *Science (New York, N.Y.)*, 318(5850):606–610, October 2007. ISSN 1095-9203. doi: 10.1126/science.1142998. PMID: 17962554.

Frank Krueger, Kevin McCabe, Jorge Moll, Nikolaus Kriegeskorte, Roland Zahn, Maren Strenziok, Armin Heinecke, and Jordan Grafman. Neural correlates of trust. *Proceedings of the National Academy of Sciences of the United States of America*, 104(50):20084–20089, December 2007. ISSN 0027-8424. doi: 10.1073/pnas.0710103104. PMID: 18056800 PMCID: PMC2148426.

Bernd Lahno. Trust, reputation, and exit in exchange relationships. *Journal of Conflict Resolution*, 39(3):495–510, September 1995. ISSN 0022-0027, 1552-8766. doi: 10.1177/0022002795039003005.

Daeyeol Lee. Neural basis of quasi-rational decision making. *Current opinion in neurobiology*, 16(2):191–198, April 2006. ISSN 0959-4388. doi: 10.1016/j.conb.2006.02.001. PMID: 16531040.

R. J. Lewicki, D. J. McAllister, and R. J. Bies. Trust and distrust: new relationships and realities. *Academy of Management Review*, 23(3):438–458, July 1998. ISSN 0363-7425, 1930-3807. doi: 10.5465/AMR.1998.926620.

J. David Lewis and Andrew Weigert. Trust as a social reality. *Social Forces*, 63(4):967–985, June 1985. ISSN 0037-7732, 1534-7605. doi: 10.1093/sf/63.4.967.

Jian Li, Samuel M. McClure, Brooks King-Casas, and P. Read Montague. Policy adjustment in a dynamic economic game. *PLoS ONE*, 1(1):e103, December 2006. doi: 10.1371/journal.pone.0000103.

Jian Li, Mauricio R. Delgado, and Elizabeth A. Phelps. How instructed knowledge modulates the neural systems of reward learning. *Proceedings of the National Academy of Sciences of the United States of America*, 108(1):55–60, January 2011. ISSN 0027-8424. doi: 10.1073/pnas.1014938108. PMID: 21173266 PMCID: PMC3017128.

Y. Long, X. Jiang, and X. Zhou. To believe or not to believe: trust choice modulates brain responses in outcome evaluation. *Neuroscience*, 200:50–58, January 2012. ISSN 03064522. doi: 10.1016/j.neuroscience.2011.10.035.

R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Dover Publications Inc., February 2005. ISBN 0486441369.

N. Luhmann. *Trust and Power*. 1979.

Babak Mahmoudi, Jose C Principe, and Justin C Sanchez. An actor-critic architecture and simulator for goal-directed brain-machine interfaces. *Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 2009:3365–3368, 2009. ISSN 1557-170X. doi: 10.1109/IEMBS.2009.5332825. PMID: 19963795.

George Mailath. Reputation effects. SSRN Scholarly Paper ID 1023658, Social Science Research Network, Rochester, NY, October 2007.

Barry R Manor and Evian Gordon. Defining the temporal threshold for ocular fixation in free-viewing visuocognitive tasks. *Journal of neuroscience methods*, 128(1-2):85–93, September 2003. ISSN 0165-0270. PMID: 12948551.

Roger C. Mayer, James H. Davis, and F. David Schoorman. An integrative model of organizational trust. *The Academy of Management Review*, 20(3):709, July 1995. ISSN 03637425. doi: 10.2307/258792.

Kevin McCabe, Daniel Houser, Lee Ryan, Vernon Smith, and Theodore Trouard. A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences*, 98(20):11832–11835, September 2001. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.211415698.

Samuel M McClure, Gregory S Berns, and P.Read Montague. Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, 38(2):339–346, April 2003. ISSN 0896-6273. doi: 10.1016/S0896-6273(03)00154-5.

Samuel M McClure, David I Laibson, George Loewenstein, and Jonathan D Cohen. Separate neural systems value immediate and delayed monetary rewards. *Science (New York, N.Y.)*, 306(5695):503–507, October 2004. ISSN 1095-9203. doi: 10.1126/science.1100907. PMID: 15486304.

David M Messick and Charles G McClintock. Motivational bases of choice in experimental games. *Journal of Experimental Social Psychology*, 4(1):1–25, January 1968. ISSN 0022-1031. doi: 10.1016/0022-1031(68)90046-2.

Alan S. Miller and Tomoko Mitamura. Are surveys on trust trustworthy? *Social Psychology Quarterly*, 66(1):62–70, March 2003. ISSN 0190-2725. doi: 10.2307/3090141. ArticleType: research-article / Full publication date: Mar., 2003 / Copyright © 2003 American Sociological Association.

E K Miller and J D Cohen. An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24:167–202, 2001. ISSN 0147-006X. doi: 10.1146/annurev.neuro.24.1.167. PMID: 11283309.

P Read Montague. Neuroeconomics: a view from neuroscience. *Functional neurology*, 22(4): 219–234, December 2007. ISSN 0393-5264. PMID: 18182129.

P Read Montague, Steven E Hyman, and Jonathan D Cohen. Computational roles for dopamine in behavioural control. *Nature*, 431(7010):760–767, October 2004. ISSN 1476-4687. doi: 10.1038/nature03015. PMID: 15483596.

P. Read Montague, Brooks King-Casas, and Jonathan D. Cohen. IMAGING VALUATION MODELS IN HUMAN CHOICE. *Annual Review of Neuroscience*, 29(1):417–448, July 2006. ISSN 0147-006X, 1545-4126. doi: 10.1146/annurev.neuro.29.051605.112903.

P.Read Montague and Gregory S. Berns. Neural economics and the biological substrates of valuation. *Neuron*, 36(2):265–284, October 2002. ISSN 0896-6273. doi: 10.1016/S0896-6273(02)00974-1.

J. S. Morris, C. D. Frith, D. I. Perrett, D. Rowland, A. W. Young, A. J. Calder, and R. J. Dolan. A differential neural response in the human amygdala to fearful and happy facial expressions. *, Published online: 31 October 1996; | doi:10.1038/383812a0*, 383(6603):812–815, October 1996. ISSN ${footerJournalISSN}. doi: 10.1038/383812a0.

Michael Naef and Juergen Schupp. Measuring trust: Experiments and surveys in contrast and combination. SSRN Scholarly Paper ID 1367375, Social Science Research Network, Rochester, NY, May 2009.

Hiroyuki Nakahara, Hideaki Itoh, Reiko Kawagoe, Yoriko Takikawa, and Okihide Hikosaka. Dopamine neurons can represent context-dependent prediction error. *Neuron*, 41(2): 269–280, January 2004. ISSN 0896-6273. doi: 10.1016/S0896-6273(03)00869-9.

B. Nooteboom. *Trust: Forms, Foundations, Functions & Failures*. Edward Elgar Publishing, 2002. ISBN 9781840645453.

J. P. O'Doherty, A. Hampton, and H. Kim. Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, 1104:35–53, April 2007. ISSN 0077-8923. doi: 10.1196/annals.1390.022.

John O'Doherty, Peter Dayan, Johannes Schultz, Ralf Deichmann, Karl Friston, and Raymond J. Dolan. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, 304(5669):452–454, April 2004. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1094285.

Paul A. Pavlou and David Gefen. Psychological contract violation in online marketplaces: Antecedents, consequences, and moderating role. *Information Systems Research*, 16(4): 372–399, December 2005. ISSN 1047-7047, 1526-5536. doi: 10.1287/isre.1050.0065.

K Luan Phan, Chandra Sekhar Sripada, Mike Angstadt, and Kevin McCabe. Reputation for reciprocity engages the brain reward center. *Proceedings of the National Academy of Sciences of the United States of America*, 107(29):13099–13104, July 2010. ISSN 1091-6490. doi: 10.1073/pnas.1008137107. PMID: 20615982.

Marios G Philiastides, Guido Biele, Niki Vavatzanidis, Philipp Kazzer, and Hauke R Heekeren. Temporal dynamics of prediction error processing during reward-based decision making. *NeuroImage*, 53(1):221–232, October 2010. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2010.05.052. PMID: 20510376.

P.J. Phillips, Hyeonjoon Moon, S.A. Rizvi, and P.J. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090 – 1104, October 2000. ISSN 0162-8828. doi: 10.1109/34.879790.

Diegon A Pizzagalli, Dietrich Lehmann, Andrew M Hendrick, Marianne Regard, Roberto D Pascual-Marqui, and Richard J Davidson. Affective judgments of faces modulate early activity (approximately 160 ms) within the fusiform gyri. *NeuroImage*, 16(3 Pt 1):663–677, July 2002. ISSN 1053-8119. PMID: 12169251.

R A Poldrack, J Clark, E J Paré-Blagoev, D Shohamy, J Creso Moyano, C Myers, and M A Gluck. Interactive memory systems in the human brain. *Nature*, 414(6863):546–550, November 2001. ISSN 0028-0836. doi: 10.1038/35107080. PMID: 11734855.

Rafael La Porta, Florencio Lopez-de Silane, Andrei Shleifer, and Robert W. Vishny. Trust in large organizations. NBER Working Paper 5864, National Bureau of Economic Research, Inc, 1996.

Wiebke Potjans, Abigail Morrison, and Markus Diesmann. A spiking neural network model of an actor-critic learning agent. *Neural computation*, 21(2):301–339, February 2009. ISSN 0899-7667. doi: 10.1162/neco.2008.08-07-593. PMID: 19196231.

Kerstin Preuschoff, Steven R. Quartz, and Peter Bossaerts. Human insula activation reflects risk prediction errors as well as risk. *The Journal of Neuroscience*, 28(11):2745–2752, March 2008. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.4286-07.2008.

RA Rescorla and AW Wagner. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In AH Black and WF Prokasy, editors, *Clas-*

sical Conditioning II: Current Research and Theory, pages 64–99. Appleton-Century-Crofts, 1972.

James K. Rilling, David A. Gutman, Thorsten R. Zeh, Giuseppe Pagnoni, Gregory S. Berns, and Clinton D. Kilts. A neural basis for social cooperation. *Neuron*, 35(2):395–405, July 2002. ISSN 0896-6273. doi: 10.1016/S0896-6273(02)00755-9.

James K Rilling, Alan G Sanfey, Jessica A Aronson, Leigh E Nystrom, and Jonathan D Cohen. The neural correlates of theory of mind within interpersonal interactions. *NeuroImage*, 22 (4):1694–1703, August 2004. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2004.04.015. PMID: 15275925.

James K Rilling, Brooks King-Casas, and Alan G Sanfey. The neurobiology of social decision-making. *Current opinion in neurobiology*, 18(2):159–165, April 2008. ISSN 0959-4388. doi: 10.1016/j.conb.2008.06.003. PMID: 18639633.

David C. Rose. *The Moral Foundation of Economic Behavior*. OUP USA, December 2011. ISBN 0199781745.

Julian Rotter. Generalized expectancies for interpersonal trust. *American Psychologist*, 26(5): 443–452, May 1971.

Julian B. Rotter. A new scale for the measurement of interpersonal trust1. *Journal of Personality*, 35(4):651–665, 1967. ISSN 1467-6494. doi: 10.1111/j.1467-6494.1967.tb01454.x.

Roy J. Lewicki. *Trust in Organizations: Frontiers of Theory and Research*. SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States, 1996. ISBN 9780803957404, 9781452243610.

G. A. Rummery and M. Niranjan. On-line q-learning using connectionist systems. Technical report, 1994.

Alan G. Sanfey, James K. Rilling, Jessica A. Aronson, Leigh E. Nystrom, and Jonathan D. Cohen. The neural basis of economic decision-making in the ultimatum game. *Science*, 300 (5626):1755–1758, June 2003. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1082976.

Tom Schönberg, Nathaniel D. Daw, Daphna Joel, and John P. O'Doherty. Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *The Journal of Neuroscience*, 27(47):12860–12867, November 2007. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.2496-07.2007.

W Schultz. Neural coding of basic reward terms of animal learning theory, game theory, microeconomics and behavioural ecology. *Current Opinion in Neurobiology*, 14(2):139–147, April 2004. ISSN 09594388. doi: 10.1016/j.conb.2004.03.017.

W Schultz, P Dayan, and P R Montague. A neural substrate of prediction and reward. *Science (New York, N.Y.)*, 275(5306):1593–1599, March 1997. ISSN 0036-8075. PMID: 9054347.

Jeffry A. Simpson. Psychological foundations of trust. *Current Directions in Psychological Science*, 16(5):264–268, October 2007. ISSN 0963-7214, 1467-8721. doi: 10.1111/j.1467-8721.2007.00517.x.

Tania Singer, Ben Seymour, John P O'Doherty, Klaas E Stephan, Raymond J Dolan, and Chris D Frith. Empathic neural responses are modulated by the perceived fairness of others. *Nature*, 439(7075):466–469, January 2006. ISSN 1476-4687. doi: 10.1038/nature04271. PMID: 16421576.

Paul Slovic. Perceived risk, trust, and democracy. *Risk Analysis*, 13(6):675–682, 1993. ISSN 1539-6924. doi: 10.1111/j.1539-6924.1993.tb01329.x.

Michael J Souza, Sarah E Donohue, and Silvia A Bunge. Controlled retrieval and selection of action-relevant knowledge mediated by partially overlapping regions in left ventrolateral prefrontal cortex. *NeuroImage*, 46(1):299–307, May 2009. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2009.01.046. PMID: 19457379.

Paula C. Stacey, Stephanie Walker, and Jean D. M. Underwood. Face processing and familiarity: Evidence from eye-movement data. *British Journal of Psychology*, 96(4):407–422, 2005. ISSN 2044-8295. doi: 10.1348/000712605X47422.

Damian A. Stanley, Peter Sokol-Hessner, Dominic S. Fareri, Michael T. Perino, Mauricio R. Delgado, Mahzarin R. Banaji, and Elizabeth A. Phelps. Race and reputation: perceived racial group trustworthiness influences the neural correlates of trust decisions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1589):744–753, March 2012. ISSN 0962-8436, 1471-2970. doi: 10.1098/rstb.2011.0300.

Leo P. Sugrue, Greg S. Corrado, and William T. Newsome. Matching behavior and the representation of value in the parietal cortex. *Science*, 304(5678):1782–1787, June 2004. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1094765.

R S Sutton and A G Barto. Toward a modern theory of adaptive networks: expectation and prediction. *Psychological review*, 88(2):135–170, March 1981. ISSN 0033-295X. PMID: 7291377.

Richard S. Sutton and Andrew G. Barto. Time-derivative models of pavlovian reinforcement. In *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, page 497–537. MIT Press, 1990.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning I: Introduction*. 1998.

Golnaz Tabibnia, Ajay B Satpute, and Matthew D Lieberman. The sunny side of fairness: preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). *Psychological science*, 19(4):339–347, April 2008. ISSN 0956-7976. doi: 10.1111/j.1467-9280.2008.02091.x. PMID: 18399886.

Edward L. Thorndike. *Animal intelligence; experimental studies, by Edward L. Thorndike.* The Macmillan company,, New York,, 1911.

Edward C. Tolman. *Cognitive Maps in Rats and Men*. American Psychological Association, 1948.

Elizabeth Tricomi, Bernard W Balleine, and John P O'Doherty. A specific role for posterior dorsolateral striatum in human habit learning. *The European journal of neuroscience*, 29(11): 2225–2232, June 2009. ISSN 1460-9568. doi: 10.1111/j.1460-9568.2009.06796.x. PMID: 19490086.

Hsing-Chen Tsai, Feng Zhang, Antoine Adamantidis, Garret D. Stuber, Antonello Bonci, Luis de Lecea, and Karl Deisseroth. Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science*, 324(5930):1080–1084, May 2009. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1168878.

Amos Tversky and Daniel Kahneman. Prospect theory: An analysis of decision under risk. Levine's Working Paper Archive 7656, David K. Levine, 1979.

Vivian V. Valentin, Anthony Dickinson, and John P. O'Doherty. Determining the neural substrates of goal-directed learning in the human brain. *The Journal of Neuroscience*, 27(15): 4019–4026, April 2007. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.0564-07.2007.

Paul A. M. Van Lange. The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology*, 77(2): 337–349, 1999. ISSN 1939-1315, 0022-3514. doi: 10.1037/0022-3514.77.2.337.

M van 't Wout and A G Sanfey. Friend or foe: the effect of implicit trustworthiness judgments in social decision-making. *Cognition*, 108(3):796–803, September 2008. ISSN 0010-0277. doi: 10.1016/j.cognition.2008.07.002. PMID: 18721917.

Janine Willis and Alexander Todorov. First impressions: making up your mind after a 100-ms exposure to a face. *Psychological science*, 17(7):592–598, July 2006. ISSN 0956-7976. doi: 10.1111/j.1467-9280.2006.01750.x. PMID: 16866745.

G Elliott Wimmer, Nathaniel D Daw, and Daphna Shohamy. Generalization of value in reinforcement learning by humans. *The European journal of neuroscience*, 35(7):1092–1104, April 2012. ISSN 1460-9568. doi: 10.1111/j.1460-9568.2012.08017.x. PMID: 22487039.

Bianca C Wittmann, Nathaniel D Daw, Ben Seymour, and Raymond J Dolan. Striatal activity underlies novelty-based choice in humans. *Neuron*, 58(6):967–973, June 2008. ISSN 1097-4199. doi: 10.1016/j.neuron.2008.04.027. PMID: 18579085.

Klaus Wunderlich, Antonio Rangel, and John P. O'Doherty. Neural computations underlying action-based decision making in the human brain. *Proceedings of the National Academy of Sciences*, 106(40):17199–17204, October 2009. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0901077106.

Klaus Wunderlich, Mkael Symmonds, Peter Bossaerts, and Raymond J. Dolan. Hedging your bets by learning reward correlations in the human brain. *Neuron*, 71(6):1141–1152, September 2011. ISSN 0896-6273. doi: 10.1016/j.neuron.2011.07.025.

Henry H Yin, Barbara J Knowlton, and Bernard W Balleine. Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *The European journal of neuroscience*, 19(1):181–189, January 2004. ISSN 0953-816X. PMID: 14750976.

Wako Yoshida and Shin Ishii. Resolution of uncertainty in prefrontal cortex. *Neuron*, 50 (5):781–789, June 2006. ISSN 0896-6273. doi: 10.1016/j.neuron.2006.05.006. PMID: 16731515.

Larry S. Zweifel, Jones G. Parker, Collin J. Lobb, Aundrea Rainwater, Valerie Z. Wall, Jonathan P. Fadok, Martin Darvas, Min J. Kim, Sheri J. Y. Mizumori, Carlos A. Paladini, Paul E. M. Phillips, and Richard D. Palmiter. Disruption of NMDAR-dependent burst firing by dopamine neurons provides selective assessment of phasic dopamine-dependent behavior. *Proceedings of the National Academy of Sciences*, 106(18):7281–7288, May 2009. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0813415106.