COMPLYING WITH NORMS.
A NEUROCOMPUTATIONAL EXPLORATION

by

Matteo Colombo

PhD in Philosophy
The University of Edinburgh
February 2012

For Andrea, with care…

Contents

Word Count: 84,200

Declaration of Own Work

I declare that this thesis has been composed by myself.

I declare that this thesis is my own work and that Chapter 7 is joint work with Aistis Stankevicius and Peggy Seriès.

I finally declare that this thesis has not been submitted for any other professional degree or qualification:

Matteo Colombo

Acknowledgements

> … 'If everybody minded their own business,' the Duchess said, in a hoarse growl, 'the world would go around a great deal faster than it does.'
>
> 'Which would *not* be an advantage,' said Alice, who felt very glad to get an opportunity of showing off a little of her knowledge. 'Just think of what work it would make with the day and night!
> You see the earth takes twenty-four hours to turn round on its axis—'

Thank you to my supervisors: Andy and Peggy.

Over the course of my doctoral studies, Andy has been a shiny inspiration to me. Smart, curious and imaginative, Andy has shown me how to think big and take chances, exploring new intellectual territories and dealing critically with issues that have the potential to impact human life. And all with grace and humility.

Three and a half years ago, I had the good fortune that Peggy accepted to supervise my PhD in philosophy by writing me "I don't know anything about social norms, but would be happy to learn :-)" Peggy has been incredibly generous and affable over these years. I am sincerely grateful for the enthusiasm, empathy, competence and patience with which she has guided me in my encounter with the workings of real computational neuroscience.

I hope that some trace of Peggy's and Andy's virtues has rubbed off on me.
It's been good fun and a great privilege to work with Peggy and Andy.

Thank you to the "Bayesian gang" in Peggy's lab.
I'm especially grateful to Aistis Stankevicius, David Reichert, Luigi Acerbi, Matty Chalk and Vincent Valton for helping me to get to grips with Bayesian and Reinforcement Learning modelling. Aistis, in particular, has been a great and tireless collaborator. Without his collaboration, probably this thesis would have one chapter less…

Very special thanks are due to those who generously read and provided valuable critical feedback on various portions of my dissertation or discussed with me some of the ideas thereof. A sincere thank you to: Aaron Seitz, Bruno Averbeck, Chiara Lisciandra, Courtney Humeny, Dave Des Roches-Dueck, Julian Kiverstein, Liz Irvine, Natalie Gold, Ray Debayoty, Rhiannon Bolos, Subramanian Ramamoorthy, Suilin Lavelle, Vincent Valton.

The usual disclaimers about residual errors and misconceptions in the thesis apply.

A number of people, instead, should be held responsible for the influence they had on my idiosyncrasies concerning norm compliance…
Part of my motivation to explore social norms is in fact biographical. As an Italian living abroad, and because of my (at times shameless) character, I sometimes happen to be involved in socially-awkward situations. This has directed my curiosity towards the nature and dynamics of social norms.

During my PhD in Edinburgh, among the people who have contributed to inform this curiosity of mine by being significantly implicated in one way or another in some of my failures to comply with social norms are: Ana, Ana Maria, Angelica, Athina, Claudio, Cormac, Courtney, Danielle, Dave & the DDR Gang, Emma, Emmett, Francesca, Georgette, Grace, Jacqui, Kirsty, Koosha, Ludovica, Miriam, Niahm, Olle, Piero, Rhiannon, Rosie, Ruth, Selina, Tony.

Thank you to them all.

Money matters. My research at Edinburgh was funded by an Engineering and Physical Sciences Research Council (EPSRC) Studentship, awarded by the School of Informatics of the University of Edinburgh, and by the British Society for Philosophy of Science (BSPS) Doctoral Award, for which I am grateful.

Family values. My heartfelt thank you goes finally to my family: to my mother, to my father and to my brother. For all their trust and loving support. Grazie!

This thesis is dedicated to the newcomer to the family: my nephew, lil’ Andrea.

> ‘Tis so,’ said the Duchess: ‘and the moral of that is—
> “Oh, ‘tis love, ‘tis love, that makes the world go round!”’
>
> ‘Somebody said,’ Alice whispered,
> ‘that it’s done by everybody minding their own business!’
>
> ‘Ah, well! It means much the same thing,’ said the Duchess, digging her sharp little chin into Alice’s shoulder as she added, ‘and the moral of *that* is—
> “Take care of the sense, and the sounds will take care of themselves.”’
> ‘How fond she is of finding morals in things!’ Alice thought to herself.

Note on Publications:

Parts of the Introduction and Chapter 1 have been accepted for publication, with support from Prof. Andy Clark and Dr. Peggy Seriès.

Parts of the Introduction have been accepted for publication as:

Colombo, M. (Forthcoming). “Constitutive Relevance and the Personal/Subpersonal Distinction” *Philosophical Psychology*.

Parts of Chapter 1 have been accepted for publication as:

Colombo, M., & Seriès, P. (Forthcoming). “Bayes in the Brain. On Bayesian Modelling in Neuroscience.” *The British Journal for Philosophy of Science*.

Complying with Norms. A Neurocomputational Exploration

Abstract The subject matter of this thesis can be summarized by a triplet of questions and answers. Showing what these questions and answers mean is, in essence, the goal of my project. The triplet goes like this:

Q: How can we make progress in our understanding of social norms and norm compliance?

A: Adopting a neurocomputational framework is one effective way to make progress in our understanding of social norms and norm compliance.

Q: What could the neurocomputational mechanism of social norm compliance be?

A: The mechanism of norm compliance probably consists of Bayesian - Reinforcement Learning algorithms implemented by activity in certain neural populations.

Q: What could information about this mechanism tell us about social norms and social norm compliance?

A: Information about this mechanism tells us that:

a_1: Social norms are uncertainty-minimizing devices.

a_2: Social norm compliance is one trick that agents employ to interact co-adaptively and smoothly in their social environment.

Most of the existing treatments of norms and norm compliance (e.g. Bicchieri 2006; Binmore 1993; Elster 1989; Gintis 2010; Lewis 1969; Pettit 1990; Sugden 1986; Ullmann-Margalit 1977) consist in what Cristina Bicchieri (2006) refers to as "rational reconstructions." A rational reconstruction of the concept of social norm "specifies in which sense one may say that norms are rational, or compliance with a norm is rational" (Ibid., pp. 10-11).

What sets my project apart from these types of treatments is that it aims, first and foremost, at providing a description of some core aspects of the mechanism of norm compliance.

The single most original idea put forth in my project is to bring an alternative explanatory framework to bear on social norm compliance. This is the framework of computational cognitive neuroscience. The chapters of this thesis describe some ways in which central issues concerning social norms can be fruitfully addressed within a neurocomputational framework.

In order to qualify and articulate the triplet above, my strategy consists firstly in laying down the beginnings of a model of the mechanism of norm compliance behaviour, and then zooming in on specific aspects of the model. Such a model, the chapters of this thesis argue, explains apparently important features of the psychology and neuroscience of norm compliance, and helps us to understand the nature of the social norms we live by.

Analytic Overview

The thesis comprises 7 chapters an introduction and a conclusion.

The ***Introduction*** is in two parts. The first states and illustrates all the main claims that are articulated and defended in the following chapters. The second explains and justifies the neurocomputational perspective adopted in the thesis.

Chapter 1 lays down the beginnings of a model of norm compliance behaviour grounded on Bayesian - Reinforcement Learning neural computation. It explains in which sense the model describes some of the core features of the mechanism of norm compliance. It argues that the neurocomputational framework adopted is more progressive than alternatives to understand the mechanism of norm compliance.

Chapter 2 provides independent reason for a neurocomputational approach to norm compliance. It argues that the explanation of paradigmatic cases of norm compliance behaviour requires the appeal to neural representations. In so doing, it explains the notion of neural representation assumed in the thesis.

Chapter 3 addresses the question of what is the representational format of the background knowledge that supports norm compliance. It argues that a structured-probabilistic approach is the more fruitful to make progress with respect to this question.

Chapter 4 focuses on moral judgement. It argues for two claims. First, some central aspects of the psychological mechanism of moral judgement can be described within the RL - Bayesian neurocomputational framework laid out in Chapter 1. Second, such neurocomputational description of moral judgement can shed new light on puzzling findings about specific patterns of moral judgement.

Chapter 5 takes up the questions whether and in which sense language is a tool that constitutes moral thinking. It argues that language is unnecessary for moral thinking, and yet language can have important effects on moral thought and behaviour.

Chapter 6 gives grounds for the claim that the emotions are not ultimate motives of norm compliance. It distinguishes between different senses of reward (and punishment) and singles out the capacity for caring as fundamental for social norm compliance.

***Chapter** 7* describes an experimental, neurocomputational project I carried out. The project asks whether and to what extent social rewards, as opposed to non-social rewards, affect our learning of social norms. The chapter provides me the opportunity to put at work some of the modelling tools and concepts used and explored in previous chapters.

The ***Conclusion*** glues all together. In light of my exploration, it reconsiders the triplet of questions and answers from which my neurocomputational journey began.

INTRODUCTION.

Topic and Method

One way to say what I am up to is by a triplet of questions and answers. Showing what these questions and answers mean is the goal of my project. The triplet goes like this:

Q: How can we make progress in our understanding of social norms and norm compliance?

A: Adopting a neurocomputational framework is one way to make progress in our understanding of social norms and norm compliance.

Q: What could the neurocomputational mechanism of social norm compliance be?

A: The mechanism of norm compliance probably consists of Bayesian - Reinforcement Learning algorithms implemented by activity in certain neural populations.

Q: What could information about this mechanism tell us about social norms and social norm compliance?

A: Information about this mechanism tells us that:

a_1: Social norms are uncertainty-minimizing devices.

a_2: Social norm compliance is one trick we employ to interact co-adaptively and smoothly in our social environment.

This question-answer triplet, in essence, is the subject of what follows.

We live in an uncertain environment, and social interaction dramatically contributes to the uncertainty underlying our environment. Although uncertainty itself does not appear to possess any normative property, social norms are technologies that respond to and manage the uncertainty of our social environment. Moral cognition arose when agents began to interact engaging in different "experiments of social living"—to use John Stuart Mill's phrase. The most successful social norms, those that are likely to survive and passed on across generations, are those that are most successful at facilitating minimization of entropy, or uncertainty, given rise by agents' interactions within the social environment.

The idea that social behaviour is bound up with minimization of uncertainty is not new. Andrew Schotter (1981) analyses institutional organizations in information theoretic terms. He focuses on economic institutions and uses the frameworks of game theory and information theory to ground the claim that "social norms and institutions are devices that give structure or order to social situations" (Schotter 1981, p. 139). Schotter believes that institutions develop out of the strategies of agents interacting with each other to solve some economic problem. The process that leads to the development of institutions is described by Schotter as a "Markovian diffusion process"—that is, as a random process whose future states are determined by its most recent state and not by the entire past—whose absorbing points—that is, whose states that are impossible to leave—correspond to stable social institutions (Ibid., Ch. 3). Absorbing points are states where expectations about the behaviour of others become self-fulfilling: belief and reality correspond perfectly in that state. This analysis is congenial to what is to follow. But my theoretical framework and my focus are unlike Schotter's. My theoretical framework is the framework of what is called *theoretical or computational cognitive neuroscience*. My focus is the *mechanism of social norm compliance*.

The framework and the approach I adopt are unlike those of a number of philosophers and social scientists working, like Schotter, on social norms within the tradition of rational choice theory (Bicchieri 2006; Binmore 1994; Elster 1989; Gintis 2010; Lewis 1969; Pettit 1990; Sugden 1986; Ullmann-Margalit 1977). I do not use game theory to analyze the creation, evolution and function of economic and social institutions. I do not start with a taxonomy that distinguishes social norms from other types of regularities. Nor do I provide a formal definition of what social norms are. My account is not intended as a conceptual analysis or as a systematization of the linguistic intuitions that people have about the word 'social norm' or about what is morally or socially (im)permissible.

Most of the existing treatments of norms and norm compliance consist in what Cristina Bicchieri (2006) calls "rational reconstructions." A rational reconstruction of the concept of social norm "specifies in which sense one may say that norms are rational, or compliance with a norm is rational" (Ibid., pp. 10-11). Rational reconstructions are not aimed at describing the processes or the mechanisms of norm compliance. Although they can yield meaningful and testable predictions, they are generally not meant to provide an "account of the real beliefs and preferences people have or of the way in which they in fact deliberate" (Ibid., p. 3). My project does not consist in a rational reconstruction of this type.

What sets my project apart is that it is meant to provide a description of some core aspects of the mechanism of norm compliance. The single most original idea put forth in my project is to bring an alternative explanatory framework to bear on social norm compliance. This is the framework of computational cognitive neuroscience. What follows describes some ways in which central issues concerning social norms and social norm compliance may be fruitfully addressed from a neurocomputational perspective.

In order to qualify and articulate the triplet above, my strategy consists in firstly laying down the beginnings of a model of some core aspects of the neurocomputational mechanisms of norm compliance behaviour, and then zooming in on specific aspects of the model. Such a model, I shall argue, explains causally relevant features of the psychology and neuroscience of norm compliance. The resolute neurocomputational perspective I am taking will lead me to cross the personal-subpersonal boundary during my exploration. Partly because of this, one may wonder what the "philosophical" contribution of my thesis is and what its "scientific" contribution is.

As long as a crisp and meaningful line can be drawn between scientific and philosophical inquiry or between a scientific and a philosophical issue, we can point to the first and third Q-As in the triplet above as the "more philosophical" since they plunge, with generality and abstraction, to the foundations of social norms by asking "How should we go about to understand social norms?" (first Q-A) and "What are social norms?" (third Q-A). The second part of this introduction, on the *framework* adopted here, will explain in what consists the neurocomputational perspective I am endorsing, and will start to shed light on these two Q-As.

More generally, without qualifying an issue as "philosophical" vs. "scientific", one may ask what my project brings to the table. In what sense is my contribution an improvement over the current state of the art on social norms and social behaviour? I maintain that by drawing on the conceptual and empirical resources of computational cognitive neuroscience we can integrate in a unifying framework the growing amount of data available about the psychology and neuroscience of social behaviour. This is likely to bring coherence to scattered issues concerning social norms, and open new possibilities for making progress in our understanding social and moral behaviour. Chapter 1 will elaborate on this point by providing further reasons for why a neurocomputational approach to social behaviour should be systematically pursued.

Here is an overview of what is to come. The Introduction does two things. First, it states the topic of this work by presenting all the major claims that will be articulated and defended in the following chapters; then it explains the neurocomputational perspective endorsed here. Chapter 1 lays down the beginnings of a neurocomputational model of norm compliance behaviour and explains in which sense the model describes some core features of the mechanism of norm compliance. Chapter 2 provides some more details on the neurocomputational account on offer. It argues that the explanation of paradigmatic cases of norm compliance behaviour requires the appeal to neural representations. Chapter 3 addresses the question of what is the representational format of the background knowledge that supports norm compliance, and what approach might be more fruitful to find it out. Chapter 4 focuses on normative judgement and argues that there is an intimate relationship between (normative) judgement and uncertainty-minimization. Chapter 5 takes up the questions whether and in which sense language is a tool that constitutes moral thinking. Chapter 6 gives grounds for the claim that emotions are not ultimate motives of norm compliance. It singles out caring as a fundamental capacity for social norm compliance. Chapter 7 describes an experimental, neurocomputational project I carried out. The project asks whether and how social rewards, as opposed to non-social rewards, affect our learning of social norms. The chapter provides me with the opportunity to put at work some of the modelling tools and concepts I used and explored in previous chapters. In the Conclusion, I reconsider the triplet of question-answer from which my journey into social norms began.

TOPIC

I begin by describing six relatively uncontroversial situations, which will help me to introduce important facts and core ideas about social norms and norm compliance. The cases described in this section are meant to be no more than intuition tweakers. In section 1.2 I use these paradigm examples to extract and elaborate general claims that constitute *explananda* for any explanatory model of norm compliance. By extracting such *explananda* from these cases, I incur an inductive risk. The risk is that the paradigm cases I rely on may turn out to be examples of features belonging to different *kinds* of phenomena. This strategy is not unusual in science and philosophy, where objects of inquiry are often put into focus only as inquiry goes along. Despite such possible preliminary conflations, I hope to show that the following six cases are really about *one* kind of phenomenon: they are all about different features of one kind of behaviour.

1. Six Tales about Norm Compliance

(A) Suppose that you live in the United States and are a blunt smoker. Smoking blunts is an increasingly popular way to consume cannabis in the United States. A blunt is a tobacco cigar hollowed and filled with cannabis. Ethnographic data suggest that blunts users are a distinct group. Typically they are male, black, older teen, into Hip Hop and living in metropolitan areas in the United States (Ream et al. 2006).

Imagine that you are a blunt smoker. Some of your friends and you have pooled money to buy cannabis and a cigar. You gather to smoke somewhere. People within your group of blunt smokers share a number of expectations. You are expected to share the blunt with other members of the group. Each person is expected to take a couple of puffs and then pass the blunt to another person. If other friends come along, they may smoke, but only if they are offered by one of those who have contributed money. They are not allowed to light the blunt. One who won't pass the blunt when expected to can be sanctioned by being called "hedgehog." One who will take more puffs than expected can be ridiculed as a "steamer" (Johnson et al. 2006).

Such shared expectations can be described as rules. These rules are enforced by an argot of social control and the risk of being shunned from the group. The majority of blunt users prefer to smoke blunts in groups. Interestingly, those who occasionally smoke blunts alone tend to replicate the group practice of taking only a few puffs and then putting the blunt out (Dunlap et al. 2006).

(B) Queues are part of our everyday lives. We queue in front of ATMs, public bathrooms, at post offices, at concerts, and so forth. Queuing is a practice with many variants and local nuances. Probably, a supermarket line in Munich is not exactly like a line to get a ticket for a football match outside the San Siro stadium in Milan, or a queue to get a drink at a pub in Edinburgh, or a line to get a railway ticket at a station in Beijing.

Suppose you are queuing to buy a train ticket at a railway station in New York City. Somebody cuts in front of you. How would you react? In one of his last works, Stanley Milgram examined the reactions of queues to intruders (Milgram et al. 1986). Milgram had confederates cut into 129 queues at railway stations and other locations in New York City. All his confederates found the mere idea of cutting into a queue emotionally taxing, and some refused to take part in the experiment. Those who took part entered the queue at between the third and fourth person. The average length of those queues was of 6 persons. The confederate stepped into the queue and faced forward while saying in a neutral tone: “Excuse me; I’d like to get in here.” On 10% of occasions queue-jumpers were physically ejected from the queue. When two intruders cut the line right in front of a person, the percentage of people who reacted by verbal objections, dirty looks or physical action was 91%. Only 5% of people, however, reacted in any way when there were two other people between them and the queue jumper. One possible explanation of these results is that people felt a unique responsibility for rejecting intruders immediately in front of them. As distance from the line intruder increases, such a feeling of unique responsibility diminishes.

(C) Suppose that you find a flyer under the windshield of your car parked outside a mall. Will you throw the flyer on the street? The answer to this question will probably depend on the state of the environment and on the behaviour of the people around you.

The social psychologist Robert Cialdini addressed the question under what conditions people litter with an experiment (Cialdini, Reno and Kallgren 1990). Like in the case you have just imagined, the experimenters gave people the opportunity to litter by placing flyers under the windshield wipers of their cars. Cialdini and his colleagues varied the state of the environment where the cars were parked. In one condition, the environment was fully littered; in a second condition, it was clean. People walking to their cars could witness a confederate who either dropped trash into the environment, picked up from the street empty cans and threw them into a bin, or simply walked through that street. Cialdini found that people threw on the street the flyers they had found on their cars more often when the environment was already littered than when the environment was clean. The most littering occurred when people saw someone else dropping rubbish into a littered environment. Of the group who saw someone else picking up and bin the litter, almost none threw the flyers away on the street.

(D) This morning, after you'd woken up and brushed your teeth, you got dressed. Did you consider being a nudist for the day? Presumably, you did not. Now, consider tipping in restaurants in North America. On an average day, approximately 10% of the people living in the United States eat at sit-down restaurants. This figure, on an average month, rises to 58% (Azar 2007a). After completing their meals, almost all of these diners add to their bills an additional payment, that is: they leave a tip. Do they consider leaving no tip? Almost all of them, presumably, do not.

Tips are not legally required. Tipping is not necessary to get good service since people leave a tip only after their meal. Diners typically don't expect to meet again the servers who waited them. So, in this case, tipping cannot be sustained by repeated two-party interaction. People in North America, it seems, typically tip thoughtlessly, that is automatically and without paying attention to what they are doing, in the same way you "thoughtlessly" got dressed this morning.

(E) Imagine that on the night of April 14, 1912 you are on the *Titanic*. The vessel is sinking, the captain issues to his officers and crew to abide by the norm WOMEN AND CHILDREN FIRST. You are neither a woman nor a child and your life is in danger, yet you may not follow your survival instinct. Imagine now that one night in

May of 1915 you are on the *Lusitania*. The ship is torpedoed by a German U-boat; your life is in danger. The captain issues the order to follow the norm WOMEN AND CHILDREN FIRST. In this case, however, you will probably follow your survival instinct and ignore the captain's orders. The number and type of passengers on both the *Titanic* and the *Lusitania* were similar. Yet, the behaviour of the people on the two vessels was different.

Women and children aboard the *Titanic* were more likely to survive than males. On the *Lusitania*, instead, young males were more likely to survive than everyone else. This opposite pattern can be explained by taking into account the fact that the *Lusitania* sank in 18 minutes whereas the *Titanic* sank more slowly in 2 hours and 40 minutes—long enough for specific social behavioural patterns to emerge. You would probably enforce the captain's orders on the *Titanic* but not on the *Lusitania* because you have time to inhibit your survival instinct and follow a specific social norm only on the *Titanic* (Frey et al. 2010).

(F) "Genie" is the pseudonym for a feral child from Los Angeles. She spent nearly all of the first thirteen years of her life in isolation, locked inside a bedroom strapped to a potty chair (Rymer 1993). When the authorities found Genie in 1970, she had one of her first interactions with other people. She had not been spoken since infancy. Genie's cognitive and social abilities were found impaired. Genie had no knowledge of the social world; she could understand a handful of words and could say only "Stopit" and "Nomore." After the first seven months of treatments at a Children's Hospital, she was prevalently oblivious to the presence of people around her, she had very little social knowledge and displayed behaviour such as spitting constantly and masturbating in the presence of other people. With the help of linguists, social workers and psychologists, after years of treatments in caring environments, Genie developed some verbal and nonverbal communication skills, she became sociable with people she was familiar with, displayed an interest in music and she learned to comply with basic social norms such as DO NOT SPIT IN PUBLIC.

1.1 Norm Compliance. Nine Features

With these examples in hand, I now describe nine *apparent* features of social norm compliance. The structure of this section is such that each heading is a specific claim related to one such feature. Any paragraphs under a given heading are intended to provide additional considerations or details to articulate the heading.

1.1.1 Norm Compliance Depends on Shared, Mutual Expectations

It seems that an essential characteristic of social norms is that they are constituted by people's mutual expectations about a certain type of behaviour in a given situation (Bicchieri 2006; Elster 2009; Pettit 1990; Sugden 1986). So, people's expectations that others don't litter in the meadows, and that others expect them not to litter in the meadows constitute a social norm against littering in the meadows. If people comply with a social norm concerning a type of behaviour, they must have certain kinds of expectations concerning that behaviour. That people share certain expectations concerning behaviour of some sort with others is one reason why social norm compliance is *social* (Elster 2009). In general we don't share expectations concerning tooth brushing. When we brush (or fail to brush) our teeth we are not complying (or failing to comply) with a *social* norm.

If social norm compliance is constitutively dependent on shared expectations, then social norms need not correspond to written rules enforced by a legal system. Whether some social norm is *also* recognized by a legal code and enforced by a legal institution does not mean that social norms *need* be codified laws. Typically, people don't receive a medal from the mayor when they comply with a norm of tipping, and they don't risk troubles with the law if they fail to leave a tip in a restaurant.

Acting upon what *seem* to be shared expectations increases the predictability of social interaction and decreases uncertainty within society. If social norms are constituted by shared expectations and such expectations "encapsulate" past experience, then social norms encapsulate past experience. Thus, they act as guides "to what to expect from the future" (Douglas 1981, p. 48). The more fully social norms are constituted by expectations, the more "they put uncertainty under control;" under the pressure of social norms, behaviour tends to acquire distinct boundaries and "disorder and confusion disappear" (Ibid.). Douglas' point on uncertainty is central to my thesis. It will be articulated in Chapters 1, 2 and 4 when I argue that to have an expectation is to have a certain representation and I explain the sense in which norms are entropy minimizing-devices.

1.1.2 Norm Compliance is Intimately Related to Punishments and Rewards

Violations of social norms typically engender attitudes like anger, contempt, blame, and punitive behaviour like avoidance, ostracism, gossip, verbal abuse and physical harm directed at the norm violator. Besides anecdotal evidence, there is a substantial body of empirical evidence from experimental economics that underwrites the claim that social norms are closely connected with punishment (Andreoni et al. 2003; Clutton-Brock and Parker 1995; Sripada 2005). This link is so intimate that some philosophers and social scientists argue that social norms are *social* partly "because they are maintained by the sanctions that others impose on norm violators" (Elster 2009, p. 197).

Other people need not intentionally impose punishments on norm violators for norm violators to incur punishment. Others may impose sanctions on norm violators, even though other people are not intentionally punishing them. To clarify the point, consider Milgram's experiment described above in (B).

One form of punishment consists in feeling negative emotions. Feeling positive emotions, on the contrary, can be rewarding. Failing to comply with norms is typically emotionally taxing. If failing to complying with a norm is emotionally taxing, then norm violators typically incur punishment. But this punishment need not depend on other people's intentionally imposing sanctions. Some of Milgram's queue jumpers reported that they felt uneasy and embarrassed at the mere idea of breaching others' expectations in a social situation. Many of them were not physically threatened in any way when they cut the queue. They recalled that they were nonetheless overwhelmed by negative emotions in jumping the queue. They incurred a form of punishment even though other people did not intentionally impose any punishment on them. It seems, then, that in general punishment and reward may partly be constitutive of social norm compliance.

The causal role of punishment seems especially weighty in comparison to rewards in giving rise to and sustaining the persistence of social norms (Andreoni et al. 2003; Sigmund et al. 2001). A type of behaviour like TAKING TWO PUFFS AND PASSING THE BLUNT AROUND can become a social norm if members of a population engage in it because they believe other members will punish them—for example by ridiculing or avoiding them, if they don't. The structure of the motivation to meet others' expectations typically comprises expectations and desires of certain kinds. On the one hand, people may anticipate such feelings as unease, shame, guilt and embarrassment at the mere idea of breaching other people's expectations. On the other, they may have the desire that others do not sanction, or think bad of them, and that others possibly think well of them.

Those who are about to jump a queue or diners who are about not to tip might feel uncomfortable and awkward anticipating other people's reaction. They anticipate that victims of norm violations—for example the waitresses and patrons who do not receive a tip when they expect to be tipped—are likely to feel anger or disgust towards the violator. Third parties—like other diners—might feel contempt at the norm violator. Being the object of these kinds of attitudes, or just assuming or anticipating being the object of these kinds of attitudes, typically makes norm violators feel shame or guilt. The anticipation of feeling ashamed, or assuming that they will be the object of a negative attitude is often sufficient—it *seems*—to move people to comply with norms.

In comparison to punishment, rewards move *some* people to comply with norms, but do not prevent *all* of those who share certain expectations from norm violation. However, reward, in *some sense*, may also be causally related to norm compliance in some situations. A type of behaviour like LEAVING A TIP TO THE

PORTER AT A HOTEL can become a norm if people believe that porters are inclined to reward them for the tip, for example with a smile or with better service.

Although the last paragraphs could have suggested that reward and punishment are essentially emotional, it should be emphasized that reward and punishment need *not* be identified with (positive and negative) emotions. In general, rewards can be understood as objects or states that make us come back for more. Punishments, conversely, can be understood as objects or states that make us *not* come back for more (Schultz 2007b). More specifically, in one sense, we *could* say that reward is something desired because of a feeling of pleasure. In this sense, leaving a tip to the porter can be rewarding because it causes positive emotions: it feels good. Because it causes positive emotions, people may tend to engage in this type of behaviour under similar circumstances in the future. In a different non-colloquial sense, which will be clarified in chapter 6, reward is something we "want" because of its perceived "incentive saliency," that is, because of its capacity to stand out from its surroundings and motivate agents to approach it, regardless of its hedonic consequences (Berridge et al. 2009). Under certain circumstances leaving a tip to the porter is something we "want" to do because of its "incentive salience," it is something likely to capture behavioural control without invariably triggering an emotional reaction. Chapter 6 articulates these claims concerning motivation, reward/punishment and emotion, and argues that the emotions are probably not the ultimate motive of norm compliance. For the moment, it is worth repeating that rewarding states or rewarding behaviour need *not* be identified with states or behaviour that engender hedonic reactions.

Let us consider one last aspect of the relationship between reward, punishment and norm compliance. What is it that makes certain behaviour *associated* with rewards or punishments? Why in general do I get punished if I fail to comply with a social norm? Why may I get rewarded if I comply with norms? There are two types of answers to these kinds of questions. The first has to do with externalities, which are secondary or unintended (positive or negative) consequences of some activity. When people urinate in the swimming pool, spit in the street, litter in the park, or use the public coffee machine without contributing anything, they are imposing negative externalities to other members of society. Hence, when people engage in such types of behaviours they may get punished because they are engaging in behaviour that is harmful to the group. Norms against behaviour such as littering, which imposes negative externalities on society, are in the public interest. Violations of such norms, therefore, provoke punishment. Analogously, behaviour that brings about positive externalities to other members of society gets rewarded.

The second type of answer involves no appeal to direct harm or benefit to members of society. People's behaving as expected is what makes certain behaviour rewarding. Failing to behave as expected is what provokes punishment. A given behaviour is rewarding, in this sense, to the extent that people are certain about what their social environment will be like when somebody engages in that behaviour. Instead, uncertainty, in some sense, will make a given social behaviour associated with punishment. One preliminary way to explain the *value of certainty* is by considering the ability of agents to make plans and engage with their environment. When agents are certain about what to expect from each other, they are in the best position to make plans and take decisions. Thus, we can say that a given behaviour is associated with rewards because people engage in that behaviour as they are expected. Conversely, for raising uncertainty, certain behaviour in social situations engenders punishments.

1.1.3 Norm Compliance is Conditional on Having the Right Kind of Representation

Social norms can be *stated* as universally quantified conditionals of the form:

For every *x*, if P*x* then M*x*

where the domain of the variable *x* is any behaviour, P specifies the property that identifies the type of behaviour and M specifies some normative property. A normative property is a property that can be ascribed with normative predicates such as 'is wrong,' 'is right,' 'is good' and so forth. Social norms can involve small or large classes of agents. For example, a social norm like CHILDREN AND WOMEN FIRST involves a class of agents larger than the class of agents involved in the social norm WOMEN FIRST. Nonetheless, both social norms can be stated as universally quantified conditionals. The fact that social norms can be stated as universally quantified conditionals, therefore, does not mean that social norms do not possess many differences of nuance or that they do not admit of exceptions.

The preference to comply with norms—for example with a norm of queuing such as FIRST COME FIRST SERVED—seems to be conditional in fact. According to Bicchieri's (2006) account, people have a preference to comply with a social norm in a situation of a certain type under the conditions that they expect others to comply with that norm in that type of situation, and they believe that others think they ought to comply with that norm in that type of situation. The cues present in a given situation are important to determine the kinds of expectations that people have in that situation, and thereby they are important to determine a preference for norm compliance.

Consider Cialdini's experiment on littering described above in (C). One way to describe those results is in terms of expectations. Cialdini and his collaborators elicited certain expectations in their subjects by manipulating the salience of the cues present in the environment. The fact that people dropped trash in the environment led subjects to expect that most people littered there. With this expectation activated, subjects were less likely to have a preference to comply with a norm against littering. Conversely, when people represented the environment as calling for an anti-littering norm—for example, when they saw another person picking up trash in an otherwise clean environment—they were more likely to have a preference to comply with a norm against littering.

How we acquire the right kinds of representations, in which sense they are "right" and what is their relationship with expectations are questions that I shall begin to answer in Chapter 1 and explore further in Chapter 2 and 4.

1.1.4 Norm Compliance Does Not Depend on a Supply of Invariant General Principles

That social norms can be described as universally quantified conditionals does not entail that people apply invariant general rules to cases when they comply with norms. Put differently, the fact that people's behaviour displays regularities when people comply with norms does not entail that people's behaviour is caused by internally represented, invariant rules when they comply with social norms.

Take the case of tipping illustrated in (D) above. Tipping varies among cultures and by service industry. People in North America tip in restaurants, but they don't tip in shoe shops. The fact that service is especially good in a restaurant may lead diners in the UK to tip the waitress or the waiter. But the same fact does not lead to the same behaviour in Japan. Waitresses and waiters in Japan would find it condescending or demeaning to receive a tip for their service. So, it seems that the features that count as cues that lead to norm compliance vary across contexts and between people. If a feature makes a given situation as one that calls for norm compliance, it does not follow that the feature always makes the same type of situation as one that calls for norm compliance.

Given a type of situation—say, having a meal at a restaurant—and a type of feature—say, good service quality at the restaurant—if the feature elicits different representations (and hence, different expectations, in a sense to be explained in Chapters 1 and 2) from one token-situation to another, then people comply with norms of tipping on a case-by-case basis, depending on the way they represent the situation. Whether a feature in a social situation counts as a cue for norm compliance, and if so, what exact role it is playing there will be sensitive to other features and to the learning trajectory of the agent in that situation.

People, in general, do not comply with social norms by applying invariant, internally represented, general rules to cases. This claim will be further motivated Chapter 3, where I explore alternative representational format of social norms and in Chapter 5, where I consider the relationship between linguistic rules and moral thought.

1.1.5 Social Norms Set the Boundaries of (In-)Appropriate Behaviour

Social norms delimit the boundaries of appropriate behaviour in many different domains of social interaction by prescribing or proscribing certain types of action. Compare these two statements:

(1) Pass the blunt after a couple of puffs.

(2) Don't jump the queue.

People's decisions in case (A), described above, are shaped by the social norm stated in (1); people's reactions in Milgram's experiment, described in (B), is shaped by the social norm stated in (2). The first statement prescribes a type of action, whereas the latter proscribes an action. (1) specifies a type of action required in a certain social context; instead (2) tells us what is forbidden under certain circumstances.

A social norm can affect people's behaviour in a population even if compliance to it is not observed. Imagine this social norm: WHOEVER FIRST MAKES A PROPOSAL THAT SOMETHING HAS TO BE DONE IS DIRECTLY RESPONSIBLE FOR MAKING SURE THE PROPOSAL IS CARRIED OUT. Imagine that students in a tutorial class have this social norm. During a seminar, those students may avoid suggesting a topic for discussion because they believe that that social norm will be followed, and, hence, they will have to prepare the talk. Nobody is violating the norm here. Everybody is avoiding it, and still the norm is guiding the students' behaviour by specifying that *if* certain conditions are satisfied, a type of behaviour is likely to follow.

The way people move when they are in certain types of situations can make visible that social norms set some boundaries for our behaviour. By complying with a norm like PASS THE BLUNT AFTER A COUPLE OF PUFFS, people's behaviour involves movements of some sort. When people comply with prescriptive norms, they behave in such a way as to form a recognizable pattern of movements. So when people tip at restaurants—thereby complying with the norm LEAVE A TIP AFTER YOUR MEAIL AT THE RESTAURANT—they typically take a look at the bill, add a certain percentage of the bill and leave the total on the table where they are sitting. Other social norms tell people *not* to move in certain ways. When people comply with such a norm as DON'T JUMP THE QUEUE they refrain from moving in certain ways: they typically stop and wait in line after the last person in the queue. This does not mean, however, that movements of some type are conceptually required to comply with norms. Norm compliance cannot be identified with recurrent patterns of movements.

1.1.6 People are Subject to Many Types of Motivations

People at any given time have multiple types of motivation. Social norms are one of such types. Social norm compliance itself has a complex motivational structure underlain by many systems as I shall explain in Chapters 1 and 6. Let's begin to consider the claim that people at any given time have multiple sources of motivations.

Social norms may have significant motivational effects on people who hold them, but they are not the only source of motivation. Frey et al.'s (2010) study presented above in (E), about the different pattern of behaviour displayed by the passengers on the *Titanic* and on the *Lusitania*, illustrates this point. Dramatic differences in behaviour may have different sources of motivation. One such source is narrow self-interest. People motivated only by self-interest are concerned about their own welfare and they don't care about other people's preferences or welfare. If people are both narrowly self-interested and instrumentally rational, then, given the state of their knowledge about the outcomes of their possible actions, they will choose action *a* if *a* is the action they believe will lead to the outcome they prefer, regardless of what *A* may involve for other people's welfare. If people are both self-interested and instrumentally rational, then they would be motivated to comply with norms only if there is some clear benefit to themselves. But human motivation is complex and does not seem to consist of narrow self-interest only.

Instrumental motives can integrate, override, inhibit, compete or interfere with other motives such as the motivation to comply with norms of cooperation or altruism. Frey et al. (2010) suggest that passengers on the *Lusitania* were mainly motivated by self-interest, whereas on the *Titanic* people complied with norms for non-instrumental, non-selfish motives. People on the *Titanic* would have followed certain norms even though there was no obvious personal benefit to them from doing so. This difference in motivation would have led to differences in behaviour aboard the sinking ships.

Complying with pro-social norms might generally take more time than behaving out of self-interest. This might be the reason why self-interest had more motivational force than pro-social motives on the *Lusitania*, which was rapidly sinking. Because on the *Lusitania*, unlike the *Titanic*, people were under extreme time pressure, self-interest might have had more motivational force than pro-social motivation there. This is consistent with one of the conclusions that Darley and Batson (1973) draw from their famous "Good Samaritan experiment."

In a nutshell, Darley and Batson found that people in a hurry are less likely to help a “shabbily dressed person slumped by the side of the road,” even if they are going to speak on the parable of the Good Samaritan. Some literally stepped over the seemingly distressed person on their way to the next building, where they had an appointment. From their results, Darley and Batson suggest that it would not be unreasonable to claim that “ethics become a luxury as the speed of our daily lives increases (Darley and Batson 1973, p. 107). A different explanation they consider is that their subjects could have been blind to the scene; that is, “because of the time pressures, they did not perceive the scene in the alley as an occasion for an ethical decision” (Ibid., p. 108).

All in all, situational forces such as time constraints seem to have a strong influence on people’s motivational dynamics.

1.1.7 Social Norms have Special Motivational Grip

Many social norms have no obvious instrumental significance. For many social norms people don’t comply with them as a means to attain some further goal, for example because of the prospects for economic gain, or future reciprocation.

There are social norms that regulate behaviour in revenge. Such social norms can motivate people to impose suffering to others who have broken a deal or dishonored a woman at some cost or risk to themselves. *Prima facie*, complying with such norms is likely to produce suffering, pointless risks and exposure to harm. Complying with a norm of revenge involves no independent benefit, that is, no benefit independent from not being punished, if you comply with that norm. At least, it is dubious whether complying with such norms can be a means to attain some further end (Elster 1990). It is reasonable to hold, therefore, that norms of revenge have powerful, non-instrumental, motivational effects on the people who have them. The same conclusion can be drawn about norms of tipping. It seems implausible to explain tipping at a highway diner that one will never visit again by appealing to instrumental rationality.

Sripada and Stich (2006) refer to the motivational grip that norms can have on people who hold them as "*intrinsic motivation.*" Their claim is that people "display an independent intrinsic source of motivation for norm compliance, and thus that people are motivated to comply with norms *over and above* (*and to a substantial degree over and above*), what would be predicted from instrumental reasons alone" (Sripada and Stich 2006, p. 285). This claim is underwritten both by the phenomenology of norms and by findings from experimental economics.

If we consider the subjective experience that often accompanies norm compliance, then it seems that many norms possess the authority to draw us to act in accordance with them unconditionally. We often don't even question the authority of the norm; we don't consider whether to comply or not. We just comply.

If we consider experimental evidence, a wealth of results shows that in variety of experimental games people comply with norms of fairness and cooperation even when that is not the most profitable thing to do (e.g. Camerer 2003). By complying with such norms, people behave very differently from the way instrumental rationality alone would predict. Once we recognize that many norms possess this type of motivational grip on people, we may want to explain the nature and origin of the intrinsic motivational power typically possessed by social norms.

1.1.8 Complying with Norms is Thoughtless

Take the case of the blunt smokers described in (A) above. There is ethnographic evidence that those who smoke blunts alone tend to behave as if they were smoking in a group, where blunts are typically consumed: they have only a couple of puffs and then put the blunt out. Or consider the situation where you enter the bank, you get in line and you wait for your turn. The behaviour displayed in both cases, it seems, is "thoughtless."

If thinking is computing, then the type of behaviour displayed in both cases requires little computation. Insofar as automatic and unconscious behaviour requires little computational effort, automatic and unconscious behaviour requires little thought. For example, we typically wait in queues without a thought, without being aware of our beliefs and preferences. Often, given certain cues, we behave automatically without conscious deliberation. Most of the time people don't comply with norms such as WAIT FOR YOUR TURN IN LINE AT THE BANK because they consciously consider that most people engage in a pattern of behaviour under that type of circumstance and that most people expect them to do the same. People, instead, tend to thoughtlessly repeat the same patterns of behaviour that they have learned both in the case of social norm compliance and, more generally, when certain situational cues trigger a determinate behaviour.

If norms put uncertainty and confusion under control—as noted with Mary Douglas above—then they spare people from a lot of computing about how to behave in social circumstances. When people are learning how to behave in social situations they are also learning "*how much to think about* how to behave" (Epstein 2001, p. 10). It seems, therefore, that another core feature of norm compliance is that "*individual thought – or computing – is inversely related to the strength of a social norm*" (Ibid.).

Enforcing norm compliance, it's important to note, might have the same feature: it may require little computational effort. If this is so, then there are grounds to resist the objection that norms cannot be sustained only by attitudes of approval or disapproval, or, more generally, only by reward and punishment. According to this objection there is always a motive *not* to enforce a norm because sanctioning of conformity and deviance is cognitively costly. Social norms—the objection goes on—can be sustained by rewards and punishments only if people have a prior, sufficiently strong, motive to maintain a system of sanctions. But being motivated to maintain such a system is also cognitively costly. Therefore social norms cannot be sustained only by people's attitudes towards certain behaviour.

This objection loses its bite if individual thought is inversely related to the strength of the motivation that people have to enforce norms by punishing norm-violators and rewarding compliers. This strength, in turn, might be directly related to the strength of the social norm that people are enforcing. Pettit (1990, pp. 739-740) makes a similar point. He notes that enforcement of norms doesn't have to involve intentional action. Intentional action may involve much thinking. If intentional action was necessary to norm enforcement, then sanctioning deviance would be cognitively costly. But the enforcement of social norms doesn't need to involve much thinking, as it need not rely on intentional action.

Pettit argues that in order to enforce norm compliance by means of rewards and punishments there is no need to go about and identify norm violators. There is no need to discipline norm violators intentionally either. It is generally sufficient that enough people are around in a social situation that calls for punishment (or reward). The simple presence of a sufficient number of other people, in fact, will be enough to (i) make it likely that the norm violator will be noticed by somebody without any active, intentional search; (ii) ensure that the norm violator will suffer some punishment without the punisher incurring any cognitive cost. It is reasonable to hold that (ii) is true: people often get punished (and rewarded) simply by believing that others think badly (or well) of them. That norm violators (or norm compliers) have this belief can be enough for them to be punished (or rewarded) without others engaging in any intentional sanctioning. "Thus—Pettit (1990, p. 741) concludes—people can be more or less involuntary enforcers of norms, automatically providing suitable rewards and punishments for acts of conformity or deviance."

Even if people's enforcement of social norms had some cognitive costs, there is evidence that enforcing social norms might be, in some sense, rewarding in itself. If the enforcement of social norms is, in some sense, rewarding in itself, then we might quickly and effortlessly overcome possible cognitive costs under the motivational pressure of the reward underlying social norm enforcement (Fehr 2009). A large number of studies in experimental economics have shown that people often punish norm violators even when the revenge brings them no personal gain, or is materially costly to them and this cost cannot be compensated in the future (Fehr and Gächter 2002). In public goods games, for example, people are willing to spend extra money to punish those who do not contribute to the public good. When non-contributors are detected, people punish them without considering whether that is in their monetary self-interest. They do it *as though* it carried the sweet psychological taste of revenge (Knutson 2004). Other experimental games have shown that also mere observers, who are not affected by others' behaviour in the game, are willing to punish others for norm violations at some cost to themselves which will not be compensated in the future (Fehr and Fischbacher 2004). This type of results suggests that people may have an "intrinsic motivation" to punish norm violators. It may be a basic feature of people's cognitive systems that the perception of a situation where a norm has been violated is sufficient to produce motivation to punish the violator. People may possess some "prior motive" to maintain a system of sanctions. I shall return on the motivational structure of norm compliance and on which sense enforcement might be rewarding in itself in Chapters 1 and 6 especially.

1.1.9 Socialization is Necessary to the Development of Norm Compliance

Social norms are found in all human societies. So, norms can be considered "cultural universals" (Sripada and Stich 2006, par. 2). This does not entail, however, that what is prescribed or proscribed by a norm is invariable across time and space or that the capacity to comply with norms is underlain by a dedicated mechanism. Different types of behaviours are proscribed or prescribed to different degrees in different groups. In the 1960's, for example, few women in specific countries wore mini-skirts, and typically they wore them only in specific situations such as in ballrooms—and then most of other people disapproved of them. Today, many women from many different places in the world wear miniskirts in a variety of circumstances—and no one gives it much thought in those places.

People of all cultures and heritages acquire the norms prevalent in their group in a reliable fashion and relatively early in life unless they already suffer some neuropsychological deficit. What seems to be necessary for the acquisition of knowledge about social norms and the development of the ability to comply with norms is socialization. The case of Genie, described above in (F), illustrates this point. Genie had serious psychopathologies and could not comply with basic social norms mainly because she had spent the first years of her life in a socially deprived environment.

That social deprivation is very likely to produce florid psychopathologies was shown by Harry Harlow in a series of controlled experiments on rhesus monkeys in the 1950s and 60s (Harlow and Harlow 1962). In Harlow's studies, the monkeys were placed in stainless-steel chambers from a few hours after birth until three, six, twelve, or forty-eighth months. The monkeys were raised with no maternal care or contact with any other living being, human or non-human; and so they couldn't develop affectional ties with their mothers or peers. When released from their isolated chambers after two years, all monkeys showed psychopathological behaviour. Two of six monkeys who had been isolated for three months stopped eating. One of these died, the other was fed with force. All monkeys behaved as if they were under extreme threat in a completely alien environment: they often assumed crouching postures with which normal monkeys typically react to extreme threat.

When paired with other monkeys, they crouched or froze; they fled when approached. They made no effort to defend themselves from assaults. Those monkeys that raised in total or partial social deprivation for more than six months had no interest in social activities such as grooming, playing and mating. They all displayed compulsive avoidance of giving or receiving emotional nourishment. Socially-deprived monkeys generally showed a specific difficulty in paying attention to other living beings. They behaved as if they could not perceive other monkeys as animal beings in their environment, and as if they were maximally uncertain as to what to do in those new circumstances.

The monkeys that could acquire approximately normal cognitive functions and display social behaviour after six months of isolation were the ones exposed to three-month old, normal monkeys (Harlow and Suomi 1971). For monkeys, and for humans alike, social deprivation is implicated in the development of important cognitive and behavioural deficits, and specifically in an inability to interact appropriately, or interact at all, with conspecifics. Social therapy consisting in interacting with others can facilitate the recovery of social capabilities impaired by being reared in isolation. This kind of study with monkeys and stories of “feral children” like Genie’s lead us to expect that socialization is essential for the development of social cognition in general, and particularly of norm compliance.

Note, finally, that the claim that socialization is necessary for the normal functioning of people’s cognitive abilities, together with the fact that social norms are a “cultural universal” (Sripada and Stich 2006) might suggest that there are innate mechanisms specifically dedicated to the acquisition and implementation of norms. Yet, current data about dynamics and connectivity of neuronal communication underlying social or moral behaviour strongly suggest that moral cognition is not identifiable through the activity of any dedicated brain sub-system (Adolphs 2010; Casebeer and Churchland 2003).

Summing up, I have described six cases (A-F) in light of which I identified nine *seemingly* core features of social norms or social norm compliance:

I. *Norm compliance depends constitutively on shared, mutual expectations.*

II. *Norm compliance is intimately related to punishments and rewards*

III. *Norm compliance is conditional on having the right kind of representations.*

IV. *Norm compliance does not depend on the application of general rules to situations.*

V. *Social norms set the boundaries of "appropriate" behaviour.*

VI. *People are subjects to many sources of motivations.*

VII. *Social norms have special motivational grip.*

VIII. *Complying with norms is thoughtless.*

IX. *Socialization is necessary for the development of norm compliance.*

It is naïve to think that there is a single, unique, simple mechanism of norm compliance. So, I don't claim that by providing a mechanistic model that could explain these features we thereby explain all there is to explain about norm compliance. Also I don't claim that an explanation of those features will provide us with necessary and sufficient conditions to identify a given behaviour as an instance of norm compliance, or to always identify the conditions under which individuals will follow a social norm.

I hold, nonetheless, that I-IX point to *seemingly* central aspects of norm compliance and can help us to develop a descriptively adequate model of one important mechanism of norm compliance. If a mechanistic model explains these aspects, then—although incomplete—it is descriptively adequate in that it accounts for a large number of observed regularities underlying norm compliance. If a model is descriptively adequate, then we have reason to believe that it has counterparts in the world and it can help us to learn about the nature of those counterparts. So when a model is descriptively adequate it can enable us to learn new things about the world. I now turn to explain the neurocomputational approach endorsed here.

FRAMEWORK

The explanatory framework adopted in this work is informed by the thesis that *the nervous system is a computing system*. This is a generic form of computationalism according to which neural computation explains cognition and behaviour. This is also one *type* of *subpersonal explanation*. I now expand on neurocomputationalism, contrast personal and subpersonal explanation and explain how neurocomputationalism can have a bearing on personal-level explanations.

1. Neurocomputationalism

Neurocomputational explanations explain how the brain carries out cognitive functions and generates behavior. They make reference to brain components—to brain areas, populations of neurons, neurons, synaptic connections, chemical neurotransmitters—and their activities, but also to the *informational* transactions between neural populations. They describe how neural processes encode, transform and decode information carried by patterns of neural activity. In some sense, which will be made clear in Chapter 2, nervous systems compute by processing neural *representations*.

Neurons' fundamental activity consists in generating all-or-none events known as spikes (or action potentials). Sequences of spikes are called neural spike trains. Depending on their biophysical properties and their connections with other neurons, neurons generate spiking trains with different properties. Neural spike trains are information-carriers and their dynamics can be described by algorithms. A *neural computation*, in the generic sense assumed here, is the transformation of neural spike trains according to an algorithm. Neurocomputational explanations consist in specifying how organized brain components and their activities produce neural spike trains that carry out cognitive functions and generate behavior (e.g. Churchland and Sejnowski 1992; Piccinini 2006; 2007).

To further clarify what I take to be neurocomputational explanation, let me introduce one of the best-developed neurocomputational explanatory models. Dopamine is a neurotransmitter implicated in many aspects of learning and decision-making. One widely accepted description of the phasic changes of activities in neurons that contain dopamine is within the framework of Reinforcement Learning (Sutton and Barto 1998). Reinforcement Learning (RL) is a field in computer science and machine learning offering a collection of algorithms to address the problem of learning what to do in the face of rewards and punishments received by taking different actions in an unfamiliar environment.

A wealth of evidence indicates that activity of dopaminergic neurons in the basal ganglia can be described as implementing a *reward prediction-error*, which is a signal used by some classes of RL-algorithms (Houk et al. 1995; Schultz et al. 1997). A reward prediction-error is the difference between obtained and expected reward. To say that dopamine neurons activity can be described as implementing a reward prediction-error is to say that some neurons can be described as performing computations by executing some RL-algorithm. By executing this algorithm, the brain would carry out the cognitive task of learning what to do in the face of expected rewards and punishments, and generate behavior accordingly. RL neurocomputation will be examined in more detail in Chapter 1. We can summarize the distinguishing features of neurocomputational explanations thus:

- *Explanatory Targets*:

Why/How does the brain carry out cognitive functions and produce behavior?

- *Explanatory Patterns*:

  Cognitive functions and behavior are explained by identifying and describing relevant mechanistic components, their organized activities, the computational routines they perform and the informational architecture of the system underlying those functions and behavior.

- *Constraints*:

  The identification of neurocomputational mechanism is constrained by spatial, temporal, structural, functional, informational and causal considerations.

- *Taxonomy*:

  The categories employed are extracted from computational cognitive neuroscience.

- *Vocabulary*:

  'Neural spiking pattern', 'Population of neurons', 'Algorithmic transformation of informational input', and the other expressions typically used to refer to the brain (or parts thereof), its activities and the computational functions it performs.

It should be clear that this sort of explanation makes no direct reference to personal-level states like beliefs and desires and to the principles of rationality that govern them. It is explanation at the *subpersonal level*. In general, explanations that deal "with parts, or systems of the cognitive agent, rather than with the agent itself as thinking and acting organism" are at the subpersonal level (Bermúdez 2005, p. 28).

Note that neurocomputational explanation is just one type of subpersonal explanation: subpersonal mechanisms can be described solely in terms of biological and chemical functions with no reference to the computational routines performed by neural activations. For example, an event like an action potential occurring at a particular time can be explained by citing distinct, antecedent events like the release of neurotransmitter molecules by a presynaptic neuron and the binding of these neurotransmitters to receptors on the postsynaptic cell. This is a case of subpersonal, non-computational explanation.

2. Personal Explanation, the Interface Problem and the Co-evolutionary Research Ideology

Explanation of people's behavior couched in the vocabulary of folk-psychology (or commonsense psychology) is instead the paradigm case of explanation at the *personal level*. The distinguishing features of this type of explanation can be summarized thus:

- *Explanatory Targets*:

  What are people doing when they behave thus and so?

  Why do people behave the way they do?

- *Explanatory Patterns:*

  - Behavior that calls for explanation is redescribed by using concepts that make it intelligible so that one now knows *what* an agent is doing in or by behaving thus and so.

  - Propositional attitudes are ascribed to agents to pick out generalizations of the form:

  "If agent S in context C desires *p* and believes that by doing *a* she will get *p*, then S will, *ceteris paribus*, do *a*."

  Such type of generalization allows us to identify the agent's *reasons* for doing *a*.

- *Constraints:*

  The ascription of propositional attitudes is based on the presumption that the agent to whom they are ascribed is rational.

- *Taxonomy:*

  The categories employed are extracted from people's everyday, "commonsense" psychological explanations, and from facts about people and their situation.

- *Vocabulary:*

  'Belief', 'Desire', 'Intention', 'Emotion', 'Reason' and other propositional-attitude expressions.

Given the distinction between personal and subpersonal level, the *interface problem* arises. The interface problem asks "how does commonsense psychological explanation [which is the prominent form of explanation at the personal level] interface with the explanations of cognition and mental operations given by scientific psychology, cognitive science, cognitive neuroscience and the other levels in the explanatory hierarchy?" (Bermúdez 2005, p. 35).

The method endorsed in this thesis makes for a *neurocomputationalist co-evolutionary* approach to the interface problem. The *co-evolutionary research ideology* is a centerpiece of the traditional neurocomputational picture of the mind (Churchland 1986). According to this position, the concepts and categories we use to understand cognition and behavior at any level of explanation "may need to be revised, and the revisionary rationales may come from research at any level" (Churchland 1993, p. 746). Hence, co-evolution involves explanations and concepts at one level being susceptible to correction and reconceptualization in light of discoveries and conceptual refinements at other levels.

From this perspective, facts about subpersonal states and events can be constitutively (or conceptually) relevant to personal-level phenomena, and therefore knowledge of such facts can, and sometimes *should*, inform personal-level explanations. This is because of one central aspect of the ordinary personal-level explanatory strategy. We ordinarily explain somebody's behavior by *redescribing* it employing different concepts. By redescribing somebody's behavior with different concepts, we make intelligible *what* someone is doing in or by behaving thus and so.

If so, explanations of phenomena like social norm compliance at the personal-level are not constitutively insulated from information yielded by knowledge of underlying subpersonal states and events; folk-psychological explanations of norm compliance in terms of preferences and expectations don't enjoy any particular autonomy from the explanations in the cognitive sciences. The concepts used in explaining human beings and their behavior can be revised under the pressure of knowledge of facts about subpersonal states and events. The extent to which this kind of knowledge will lead to a revision of the folk-psychological concepts we use to explain personal-level phenomena depends on the proper identification of neurocomputational mechanisms. If we are to understand how facts about subpersonal states and events may lead to conceptual revisions of personal-level phenomena, we must attend to the distinctive details of the neurocomputational mechanisms we identify. This is one of the burdens of this work. In the conclusion, in light of our neurocomputational journey into social norm compliance, I shall put forth some hypotheses about concepts we could use to describe this personal-level phenomenon.

It should be emphasized, finally, that adopting a neurocomputational perspective to the interface problem doesn't entail an eliminativist stance toward folk-psychology. Some advocates of a neurocomputationalist approach to the mind have put emphasis on its discontinuities with folk psychology, thereby arguing that folk-psychology is radically false and should be replaced with explanations couched in terms of our best scientific theories of how the brain works (Churchland 1981; 1995). Others like Clark (1989) argue for ecumenical views whereby folk-psychology and neurocomputational approaches to the mind have distinctly different explanatory roles, and so can peacefully coexist. Rather than hostility to folk-psychology, what motivates neurocomputationalism is a *co-evolutionary* conception of the relationship between different explanatory levels and frameworks.

CHAPTER 1.

The Building Blocks of Norm-Hungriness

This chapter describes and defends the beginning of a neurocomputational mechanistic model of social norm compliance. The workings of this mechanism can plausibly explain central features of social norm compliance. More precisely, the chapter identifies and describes putative neurocomputational building blocks of social norm compliance. In order to identify these building blocks, it firstly identifies two computational problems, which social cognition must solve to enable cognitive agents to comply with social norms. The following argument offers one way to identify the nature of such computational problems.

P1: Adaptive behaviour demands "uncertainty" minimization.

P2: Social norm compliance is an instance of adaptive behaviour.

Therefore, C: Social norm compliance demands "uncertainty" minimization.

The first part of the chapter explains each of the premises and the conclusion of this argument. In particular, it articulates the relevant notion of 'uncertainty' minimization. The second part of the chapter draws upon these explanations to describe what can be called *neurocomputational building blocks* of social norm compliance.

The first part comprises two sections. Section 1 explains the claim that adaptive behaviour demands "uncertainty" minimization (P1), by introducing Karl Friston's "free-energy" principle of adaptive behaviour. Section 2 claims that when agents comply with social norms they thereby behave co-adaptively (P2). In making this claim, it identifies two computational problems social cognition must solve to enable adaptive agents to comply with social norms, namely:

(i) To use sensory information to compute representations of social situations.

(ii) To consume these representations to determine future movements or internal changes in the presence of, and interaction with other agents.

The second part comprises three sections. It is suggested that a mechanism consisting of Bayesian-Reinforcement Learning systems can solve these problems. This suggestion leverages recent advances in (a) neural models of Bayesian inference and (b) Reinforcement Learning algorithmic accounts of how neural activity can enable learning and decision-making. By minimizing *prediction-errors*, this mechanism enables people to acquire and act upon social norms. On my account, the Bayesian system yields social representations, and the Reinforcement Learning system draws on social representations to generate actions so as to minimize *reward prediction-error* during social interaction. Sections 3 and 4 describe in detail the two systems comprised by this mechanism, and explain how they could ground norm compliance. Section 5 concludes by laying out three main arguments for why norm compliance is best understood within this Bayesian-Reinforcement Learning model.

PART I. Social Norm Compliance and Uncertainty Minimization

1. Adaptive Behaviour and Uncertainty: A Free-Energy Principle

What does it mean that adaptive behaviour demands "uncertainty" minimization? One way to address this question is by considering one recent proposal articulated by Karl Friston, which purports to connect and explain in a single unifying framework adaptive biological processes, brain function, and the relationships between cognitive functions such as action, perception and learning (Friston 2005, 2009, 2010; Friston and Stephan 2007). *Uncertainty minimization* is the fundamental notion in Friston's framework. So, by introducing and explaining the main tenets of Friston's framework, I hope to clarify and motivate the claim that adaptive behaviour demands uncertainty minimization.

A couple of caveats before I introduce Friston's proposal. Friston's theses are both controversial and interesting, partly because of the dramatic claims made for their explanatory power. I am not interested, however, in providing a critical evaluation of his proposal here (see e.g. Fiorillo 2010; Thornton 2010). Furthermore, the claim that brain function and adaptive behaviour are intimately related to "uncertainty" is not new (Dayan et al. 1995; Rao et al. 2002; von Helmholtz 1925). My choice of explaining premise P1 in the argument above from the angle of Friston's proposal depends on its generality and its explicit reference to adaptive behaviour and self-organizing systems.

According to Friston, adaptive behaviour and the structure and function of the brain can be explained "starting from the very fact that we exist" by appealing to a "free-energy" principle (Friston 2009, p. 293). "The free-energy principle says that any self-organizing system that is at equilibrium with its environment must minimize its free energy. The principle is essentially a mathematical formulation of how adaptive systems (that is, biological agents, like animals or brains) resist a natural tendency to disorder" (Friston 2010, p. 127). Let me elaborate.

Friston starts from the fact that homeostatic processes ground life. All biological, adaptive, self-organizing agents resist a tendency to disorder by maintaining their state and gross form in the face of a constantly changing environment. All adaptive agents, that is, possess homeostatic properties which enable them to maintain their physiological and sensory state within bounds. Without these properties, life would not be viable. Warm-blooded animals, for example, could not exist without their homeostatic properties, which maintain their temperature within a certain range. The repertoire of physiological and sensory states in which adaptive, biological systems can be is limited. If a biological system is in some physiological or sensory state outside certain bounds, its homeostatic relations will break down and it will soon die. Friston restates this fact by employing mathematical tools and notions from information theory.

The key concepts of his framework are the information-theoretic notions of *entropy*, *surprise* and *free-energy*. They are all intimately related to the notion of *uncertainty*, as information theory is precisely the branch of mathematics that describes how uncertainty should be quantified, manipulated and represented. Information for a system consists in the reduction of uncertainty for that system. So the uncertainty of a system decreases as it receives information (Shannon 1948; see MacKay 2003 for an advanced textbook treatment of information theory). Let me now introduce each notion, and explain how such concepts could bear on explaining adaptive behaviour and brain function.

If we describe with a probability distribution all possible physiological and sensory states in which an adaptive agent can possibly be, then the point that adaptive agents must resist a tendency to disorder can be re-stated by saying that that distribution must have low *entropy*. Entropy, in information theory, measures the amount of uncertainty of a random quantity. That a probability distribution has low entropy means that the outcomes sampled from that distribution are relatively predictable. Conversely, outcomes sampled from distributions with high entropy are relatively unpredictable. If the probability distribution of the possible sensory and physiological states of an adaptive agent has low entropy, then the agent will occupy relatively predictable states.

One way to rigorously characterize the informal notion of a predictable state, or outcome is in terms of the amount of *surprise* (or *surprisal*) associated with that state. Surprise quantifies how much information an outcome carries for a system. The amount of surprise of a particular outcome is a function of the probability of that outcome, such that the less probable the outcome the more surprising the outcome is. The amount of surprise of two independent outcomes is the sum of the amount of surprise of each outcome. Given these two properties, the surprise of an outcome should be the negative log-probability of that outcome. Entropy of a probability distribution is just the average amount of surprise of outcomes sampled from it.

It should be clear that entropy and surprise are measures relative to a probability distribution, or an agent. For example, you may have high uncertainty about the result of the match tomorrow, but your teammate may not. This results in different entropies, or surprises associated to the same outcome (i.e. the result of the match). More on this point in a moment.

Now, suppose that we describe with a probability distribution all possible physiological and sensory states in which a fish can possibly be. An improbable outcome from such a probability distribution is “fish out of water.” Because it is improbable, this outcome is surprising for the fish. Conversely, a probable outcome from that probability distribution is “fish in water.” This outcome is not surprising for the fish. Because surprising outcomes are those that correspond to a likely breakdown of the homeostatic relational properties of the fish, the fish must avoid surprising states in order to have highest probability to exist and keep on existing. The probability distribution describing its (viable) sensory and physiological states must have low entropy. Biological agents ensure that their sensory entropy remains low and that they live longer by minimizing the long-term average of surprise of the probability distribution describing all their possible states.

Three points are worth emphasizing. First, as mentioned above, entropy and surprise can only be defined in relation to an agent (or a probabilistic model). When applied to adaptive agents, average surprise, or entropy is a function of a sensory state and the agent’s internal model of the environmental causes of its sensory state. Agents could be thought of as maintaining internal, probabilistic models of the relevant variables in their environment causing their sensory states. These models are tuned by learning and experience, as the agent interacts with its environment. An agent’s model can be understood as corresponding to the agent’s uncertain “beliefs.” The next Chapter will distinguish between different senses of “belief” in terms of explicit, implicit, tacit, conscious and unconscious representations. For the moment, suffices to say that “belief” here does not necessarily refer to an introspectively accessible or conscious mental state. Rather it corresponds to an “implicit” probabilistic internal representation, which affects the agent's behaviour. It should be clear then that in function of an agent's internal model, the same state can be surprising for one agent but not for another agent. Even for the same agent, the same state may carry different amount of surprise at different times. I shall return on "internal models" below in this chapter in relation to how agents acquire social representation through the workings of their Bayesian brains.

The second point is that "surprise," in the context of Friston's framework, should be distinguished from the subjective point of view of a conscious agent. The two notions are distinct. Avoiding subjective surprise need not imply avoiding surprise in the information-theoretic sense. For example, you may consciously judge that you are in a surprising situation if you perceive that your cat Piper is speaking to you. However, if Friston is right, and cognitive, biological agents are mandated to minimize the uncertainty of their sensory and physiological state over their lifetime, then this percept is the one that most effectively minimizes the long-term average of surprise (or entropy) of your sensory states—regardless of your subjective, conscious judgement.

Third, it can be considered a tautology to say that agents that are in unsurprising states are in those states frequently and they keep existing by being in those states. It would amount to a re-description in information-theoretic terms of one aspect of biological systems that exist. Information theory provides us with one possible quantitative framework whereby we can describe adaptive behaviour and cognition, but it is not clear how this would explain or provide special understanding on adaptive processes and cognitive phenomena. In other words, what is it that Friston brings to the table?

Friston proposes a principle, which could explain how surprise minimization is carried out by computationally-bounded cognitive systems. Computationally-bounded agents cannot evaluate and minimize surprise directly since this would entail that they "know" all the variables of the world causing their possible sensory states. Adaptive, computationally-bounded agents are proposed to minimize "free-energy" instead, which is a quantity that provides an upper bound on surprise *and* can be directly evaluated and minimized by computationally-bounded agents.

Free-energy, as characterized by Friston, is an information-theoretic measure "that bounds or limits (by being greater than) the surprise on sampling some data given a generative model," where a generative model describes a process assumed to give rise to some data (Friston 2009, p. 209; Friston 2010, p. 127). In this context, a generative model is defined in terms of both a prior distribution over the environmental causes of an agent's sensory states and the generative distribution (or likelihood) of the agent's sensory states, given the environmental causes of those states. The generative model generates sensory states from their causes.

Friston shows that free-energy provides a bound on surprise. So, to the extent that the bound is tight, minimizing free-energy minimizes the probability that agents occupy surprising states. Minimization of surprise via minimization of free-energy is a feasible process. The free-energy of an agent would depend only on its sensory states and its internal model of the environmental causes of its sensory states. Since both sensory state and internal model can be evaluated and manipulated by computationally-bounded agents, free-energy can be directly minimized by computationally-bounded agents. Since the free-energy of an agent is a function of the internal state of the agent's brain, which embodies a model of relevant environmental variables, and of sensory data, free-energy minimization would provide a mechanism by which adaptive agents can avoid surprising states and thereby live longer.

Summing up. I have explained the claim that adaptive behaviour demands uncertainty minimization by introducing the information-theoretic notions underlying Karl Friston's "free-energy" framework. The specific value of Friston's hypothesis is controversial and I have not tried to give a critical assessment of it. What I have hoped to have clarified is that we can view the process by which agents adapt and interact successfully with their environment as a process by which they reduce their uncertainty. Now I illustrate this claim by focusing on a case study and on the notion of *prediction-error*.

1.1 Learning to Play and Uncertainty. Prediction-Error Signals

TD-Gammon is a neural network that is able to achieve master level skills at the game of backgammon (Tesauro 1994; 1995). In backgammon two players take turn rolling a pair of dice. Each player has 15 pieces which can be moved on a board of 24 locations. The roll of dice determines how far players can move their pieces. The first player to remove all of her pieces from the board wins. Backgammon is highly stochastic and good play requires strategic skills. With each roll of the dice, players have to choose from numerous options for moving their pieces. The pieces can interact as they pass each other going in different directions, and so players ought to anticipate possible moves by the opponent. Although the number of possible backgammon configurations is enormous, a complete description of the state of the game is available at all times and is given by the configuration of the board. The outcome of the game is easily identifiable and can be treated as a final reward to be predicted.

TD-Gammon uses a standard multilayer perceptron (MLP) architecture (Figure 1).

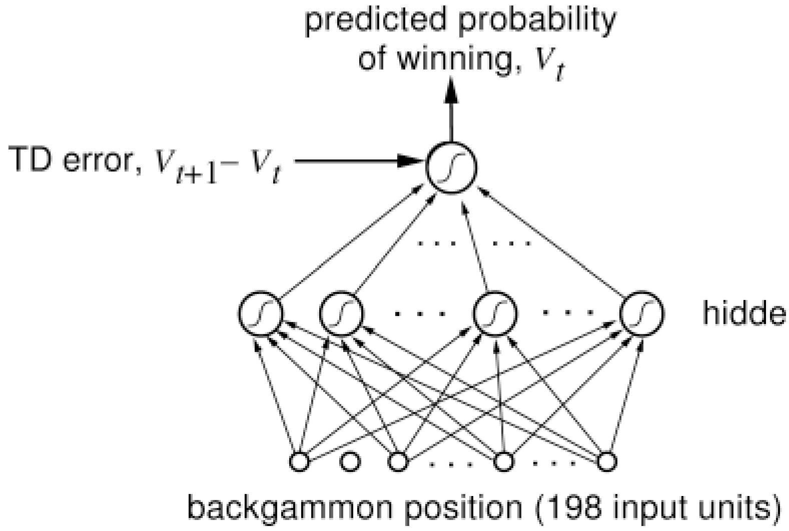

predicted probability of winning, V_t
TD error, $V_{t+1} - V_t$
hidden units (40-80)
backgammon position (198 input units)

Figure 1. TD-Gammon: an artificial neural network trained by a form of temporal-difference learning. (From Sutton & Barto 1998, Figure11.2)

It has a layer of input units, a layer of hidden units and a layer yielding outputs. Each of the connections between units is parameterized by a real valued weight. The weights embody the network's strategic knowledge of the game. The input to the network is a representation of a backgammon board configuration. For each input pattern, TD-Gammon yields an output vector indicating the predicted probability of winning the game. One strategy TD-Gammon can use to improve its game is to learn to make accurate predictions. One way to learn to make accurate prediction is by means of a reward prediction-error. To say that a system minimizes prediction-error is another way to say that it minimizes its uncertainty.

TD-Gammon started with weights set at random, and so it had no knowledge about how to play good backgammon. It was trained by self-play: the same network chose the moves of two opposite players during training. At each time step, which corresponded to some move made by one side, TD-Gammon executed a non-linear form of temporal difference (TD) learning algorithm to minimize reward prediction-error and change its weights. By executing the TD-algorithm the network learned a value function *V* that evaluated board configurations *S*. At each time step *t*, the network acquired a representation of the board state s_t in the game. From this representation, it produced a number V (s_t; $\mathbf{w}$) which specified how good the state represented was. The weights **w** of the network were tuned during learning so that the evaluation function *V* could accurately describe the probability of winning the game moving from configuration s_t.

After a million games, TD-Gammon's knowledge of how to play improved to the extent that it could play on a par with the best human players. The key of TD-Gammon's success is the reinforcement learning algorithm mentioned above: the temporal difference reward prediction-error, which we also encountered in the Introduction.

Reinforcement Learning (RL) studies the ways natural and artificial agents can learn to predict the consequences of their behaviour and optimize it in environments where actions lead from one state to the next and can lead to rewards and punishments. TD-Gammon illustrates a fundamental insight of RL-models: how agents can develop intelligence and flexible behaviour by interacting with other agents and their own environment. Prediction-error minimization is the engine of such processes.

In general, the prediction-error approach consists in using past knowledge and current experience to predict what the future holds. A prediction-error is a difference between an actual and an expected outcome. This discrepancy is used to update expectations in order to make predictions more accurate. A *reward* prediction-error is a difference between two values associated with executing actions in some state. The *value* of a state is the expected sum of future rewards and punishments that can be achieved starting to act from that state. In its most simple form, the reward prediction error δ_t is the difference between the predicted value (V_{t+1}) and the current value (V_t) of a given state at time t:

[1] $\delta_t = V_{t+1} - V_t$

Equation [1] is foundational to many models in cognitive science, from conditioning models (Rescorla and Wagner 1972) to more elaborate connectionist models of cognition and learning (Rumelhart, McClelland, and the PDP Research Group 1986) to the most recent models of brain function (Dayan and Abbott 2001; Niv and Schoenbaum 2008). Friston himself claims that the free-energy of a system "is just the amount of prediction error" in the system (Friston 2009, p. 293).

Let us now ask: How did reward prediction-error minimization enable TD-Gammon to play at grandmaster level? To reach expert play, TD-Gammon learned the value of various positions on the board in terms of the probability to win the game moving from that configuration. It learned these values by adjusting its weights in function of its predictions-errors: in function of the discrepancy between its predictions before and after a move. Given a board configuration s_t, TD-Gammon predicted the probability of winning before making a move from that board position. The move selected at each time step was the move with highest probability of winning the game. TD-Gammon observed the actual outcome. If its prediction was wrong, the system would update the knowledge-base embodied in its weights to make its predictions more accurate. As the predictions became more and more accurate with experience, information about the value of making a certain move from a given position propagated towards the earlier stages of a game. If its predictions were correct, no prediction-error would have occurred because the prediction based on the configuration at time t would have been equal to the predicted outcome from one time step later t+1 and onward. So, by minimizing prediction-error, TD-Gammon learned the objective probabilities of winning the game starting from a given position. Put differently, by minimizing prediction-error, TD-Gammon built a map of objective values for each of the possible configurations on the board. In order to be useful, however, this map should be able to influence behaviour that preempts the consequences of decisions. How could this happen?

Part of the answer is in the functional significance of error signals. Action selection can in fact be driven by error signals since they convey information about whether a certain action led the agent to a state with higher value than the previous state—recall that a state with higher value is a state predictive of more future reward (Sutton and Barto 1998, Ch. 6.6). If the prediction-error is positive, then the chosen action led the agent to a "better" state, for example a board configuration that improved the prospect of winning. Given the goals of the agent, for example winning the game, the tendency to select that action should be strengthened for the future. A negative prediction-error signals that the tendency to choose that action should be weakened since it brought about a state "worse" than the previous one. Thus, the agent can build an action plan (or *decision policy*) π (*s*, *a*), according to which the probability *Prob* ($a \mid s$) to perform certain actions *a* at each state *s* is increased or decreased based on the error signal that follows each action:

$$[2] \qquad \pi (s, a)_{\text{new}} = \pi (s, a)_{\text{old}} + \eta_{\pi}\delta_{\text{t}}$$

where η_{π} is the learning rate of the action plan and δ_{t} is the prediction error at time step t. Reward prediction-error minimization, therefore, can lead an agent both to build an accurate map of the predictive value of each state and to select the action that leads to a better state in its environment.

The prediction-error approach may be generalized in a number of ways. I shall argue that it can be fruitfully extended to the domain of social behaviour. Now I turn to explaining the claim that when agents comply with social norms they thereby behave (co)adaptively (which is premise P2 in the argument at the beginning of this chapter). I identify and explain two problems of prediction-error minimization that we face in our social world. By solving these problems, agents satisfy—at least partly—the demands for uncertainty minimization posed by social norm compliance (as stated in the conclusion C of the argument above).

2. Social Brains and Uncertainty.
Two Computational Problems for Social Cognition

Human agents live in a world populated by other people. We are bound to act in the presence of others. But we are also bound to *interact* with others. Human beings are essentially social animals. Our interaction with others and the relationships we form with other people are enormously important to us, both for our material life and for our cognitive functioning.

We need to interact with others to fulfill most of our material needs. In general, the ways we get to live in a house, acquire food and most of other material goods depend on social interactions. Without interacting with others it would be extremely difficult, if possible at all, to satisfy our basic needs for food and shelter. Our normal cognitive development, moreover, is dependent on being exposed to social stimuli and on caring relationships. Recall the cases of "Genie" and Harlow's monkeys described in the Introduction. They illustrate that solitary confinement is the most significant cause of many psychopathological conditions, which can in fact be effectively treated with the help of social therapy and caring relationships. Some also argue that our cognitive flourishing is itself a socially embedded process (Doris and Nichols Forthcoming). If cognitive flourishing is a socially embedded process, then optimal cognitive functioning causally depends on and is sustained by sociality. Insofar as our cognitive flourishing is sustained by social interaction, we best judge and make good decisions when our judgements and decisions are part of a social process. Yet, it should be clear that the fact that we are social creatures and sociality is so important to us does not entail that interacting with others does not pose demanding computational challenges to social cognition. Rather, the importance of sociality makes these challenges more pressing.

The major computational challenges faced by social cognition are two:

(i) To use sensory information to recognize, that is, to compute representations of, social situations.

(ii) To consume these representations to determine future movements, or internal changes, in the presence of and interaction with other people.

In a sense, any cognitive system needs to find some way to use sensory representations to determine bodily changes or future movements so that it can behave adaptively in the world. In a sense, then, (i) and (ii) are not computational problems specific to social cognition. In the domain of social interaction, however, they are much more complicated because living with others makes our surroundings more uncertain, complex, noisy and ambiguous.

If challenges (i) and (ii) are not specific to social cognition, then reliable computational solutions for perception and motor control might be extended to the domain of social interaction. One such solution is prediction-error minimization. Prediction-error minimization in a social environment can facilitate agents to adapt their behaviour to each other's, and thereby to interact smoothly. Let me start to unpack what I mean by 'co-adaptation' and why it is important for sociality.

The behaviour of two or more agents is *co-adaptive* if it contributes to the agents' satisfying their desires, preferences and needs in the environment in which they are embedded. Agents are best able to make plans and satisfy their desires when they are able to predict what their environment will be like over time. Since human agents are embedded in a social environment, they are best able to make plans and satisfy their desires when they are able to predict each other's behaviour and changes in their social landscape. It is easier to make plans and satisfy one's desires when we are surrounded by agents who routinely engage in "normal," expected behaviours. By acting on such predictions about each other's behaviour, agents can adjust their behaviour to each other's. When agents adjust their behaviour to each other's in this way, their predictions about each other become self-fulfilling, and thereby they can deal with their surroundings more intelligently and at little computational cost. If we are bound to share the environment and interact with our conspecifics, then people’s behaviour must co-adapt to each others’ behaviour. When people behave intelligently and adaptively in their social world, they thereby generally interact *smoothly* with each other.

Smooth interaction with others is a process involving fluid, thoughtless, context-sensitive responses to incoming social stimuli. If, for each of our social interactions, we had to negotiate every decision we make by inquiring other people about their needs, their desires, their beliefs, their entitlements and so on, we would spend most of the time engaged in effortful, time-consuming thinking. We would not get much accomplished. We would not have even time to engage actively with others: we would just ponder about people rather than act with them.

If, for example, we always tried to figure out the distance we should keep from each of the people we meet to make them most comfortable, the flow of our interactions would be continuously interrupted and we would undergo massive cognitive costs. The number of parameters we would need to take into account to solve such a trivial problem would be enormous. In order to compute the right interpersonal distance for every person we may meet, we would need to sample people on the street and identify the right values of parameters such as gender, nationality, personal character, social context, and so forth. This kind of thinking would hardly facilitate us to navigate our social world. We would be occupied by unimportant activities like sampling people to find out what is their right interpersonal distance. This would prevent us from engaging in activities necessary for our material well-being and cognitive flourishing that require some cognitive load. Smooth interaction seems, therefore, to be necessary to navigate intelligently our social surroundings so that we can more easily get those things that we regard as important to ourselves accomplished.

The prediction-error minimization approach can be used to solve the two problems stated above:

(i) To use sensory information to compute representations of social situations.

(ii) To consume these representations to determine future movements or internal changes in the presence of, and interaction with other agents.

By meeting these two challenges, prediction-error minimization enables people to acquire and act upon social norms. Acting upon social norms facilitates people to adapt their behaviour to each other's, and contributes to make our social interactions smooth. The types of prediction errors being minimized to solve those challenges are three:

- a *sensory input* prediction-error,
- a *reward* prediction-error and
- a *state* prediction-error.

The first type of prediction error helps agents to solve challenge (i); the last two types to solve challenge (ii). TD-Gammon illustrates how the reward prediction-error signal can be used to learn values for action choices that maximize expected future reward. Sensory input prediction-errors report discrepancies between the expected and the current sensory input. The next sections focus on these three types of prediction-error explaining how, by solving challenges (i) and (ii), the systems that generate them are building blocks of norm compliance behaviour.

PART II. Towards a Neurocomputational Model of Social Norm Compliance

3. Social Representations and Bayesian Brains

How do we acquire social representations? By minimizing reward prediction-error TD-Gammon came to “see,” at least indirectly, features in its world to which it was previously insensitive. TD-Gammon acquired a kind of perceptual skill that enabled it to play more and more proficiently. I wish to show in this section that people acquire a similar perceptual skill courtesy of the computation of richer and richer representations of their social situation. The computations of social representations are carried out by means of Bayesian inference.

The next Chapter focuses on the topic of neural representations and characterizes what neural representations could be and why we need them to explain norm compliance. To a first approximation, here a *representation* is understood as a neural event that carries information about some state or situation in the world. For now, by saying that some neural event *encodes* some representation I mean that some neurons or populations of neurons are the vehicles of some piece of information about some state in the world.

In general, a *state* (or a situation) is a set of variables in a process that generate sensory data or inputs. States, that is, cause sensory inputs. Typically such variables vary rapidly and continuously over time. Hence, states of the world change rapidly and continuously over time. In processes generating sensory inputs, some variables, however, change discretely and on a slower time scale. Sets of these discrete and slowly changing variables can be called *contexts*. The distinction between state and context is important, but for my argument it won’t make a significant difference. For ease of discussion, if not otherwise specified, I use ‘state’ to refer to both states and contexts.

States in the environment stand in causal relationships between each other. Such causal relationships can be referred to as *structure*. Different relationships between states—different structures, that is—can be expressed in mathematical equations and depicted by means of graphical models (Vilares and Kording 2011).

The only access we have to the world is through our senses which can be viewed as sources of information about the states of the world and their structure. This information is generally corrupted by random fluctuations, noise and ambiguity. The same sensory information can be caused by many different states and the same states may cause different types of sensory information. When we act in the world, moreover, our motor signals are also corrupted by noise. Since intelligent and adaptive behaviour is tied to the ability to survive in a changing and uncertain environment, our cognitive system must handle sensory and motor uncertainty in order to extract information about which state obtains in the world. The Bayesian framework provides one principled way this sensory and motor uncertainty can be handled in order for us to behave adaptively in our world.

Bayesian inference is a type of statistical inference where data (or new information) are used to update the probability that a hypothesis is true. To say that a system performs Bayesian inference is to say that it updates the probability that a hypothesis H is true given some data D by executing Bayes’ rule:

[3] $\mathit{Prob}\ (\mathrm{H}|\mathrm{D}) = \mathit{Prob}\ (\mathrm{D}|\mathrm{H})\mathit{Prob}\ (\mathrm{H}) / \mathit{Prob}\ (\mathrm{D})$

We can read [3] thus: “the probability of the hypothesis given the data ($P(H|D)$) is the probability of the data given the hypothesis ($P(D|H)$) times the prior probability of the hypothesis (*P*(*H*)) divided by the probability of the data (*P*(*D*))." In the case of our cognitive system, hypotheses can consist of either structures or states of the world, and data correspond to sensory inputs. As our cognitive system receives sensory information, the probability distribution over the possible structures or states of the world is updated via [3].

Our cognitive system can be described as having top-down and bottom-up signals. Top-town signals represent prior expectations about states in the world before we receive sensory information, formally:

Prob (State).

Bottom-up signals represent sensory information conditional on prior expectations, formally:

Prob (Sensory Input | State).

When the bottom-up signal does not make any difference to our cognitive system, then our expectations about the states in the world remain unchanged. No sensory prediction-error is generated. When the bottom-up signal makes a difference, then, by multiplying the prior by the likelihood and normalizing, our cognitive system can compute the posterior probability:

Prob (State | Sensory Input).

This posterior, in turn, becomes the new prior about states obtaining in the world and can be further updated based on new sensory input. The execution of this updating is carried out by what can be called *a sensory input prediction-error*. If these errors in sensory prediction are systematically translated into changes in synaptic weights, then we would have a Bayesian neurocomputational mechanism of perception. The

Bayesian framework will be further discussed and put at work in Chapter 4 in relation to moral judgement.

The problem of how, given a *structure* in the sense above, our cognitive system can infer the hidden cause that generated sensory input is currently a major topic of research in computational neuroscience. There are accumulating pieces of evidence that indicate that the cortical network might implement Bayesian inference (Doya et al. 2007; Knill and Richards 1996; Rao et al. 2002). There are three sources of evidence. The most telling comes from psychophysical experiments where people's performance is shown to approximate the Bayesian optimum. Besides psychophysical experiments, a number of computational models show how approximate Bayesian inference could be implemented in biologically plausible neural networks. Finally, broad features of biological sensory systems can be explained in a Bayesian framework. Let me expand on this last point.

Sensory processing takes place along a cascade of many processing stages over cortical areas arranged in a hierarchical structure. This basic structural feature would be explained by Hierarchical Bayesian models of sensory processing where Bayesian transformations are temporally sparse, with processing time scales getting progressively longer as one moves up the layers, and spatially distributed along multiple layers of a hierarchy (Lee and Mumford 2003). Moreover, the anatomy and physiology of inter-regional connections in the cortical hierarchy point to a functional asymmetry between forward and backward connections. Forward connections run from lower to higher cortical layers and seem to drive neural responses. Backward connections run from higher to lower layers and mainly play a modulatory role by affecting neural responsiveness to other inputs. This functional asymmetry can also be explained within Hierarchical Bayesian models of sensory processing. According to one model (Friston 2008), cortical hierarchies generate sensory data from representations of causes at high-levels. Thus, prior knowledge about the causal structure of the environment would be encoded in the backward connections. Forward connections would provide feedback by transmitting sensory prediction-error up to higher levels. Perception would arise from mutually informed top-down and bottom-up transformations distributed along the hierarchy.

3.1 Bayesian Computing of Social Representations

What could *social* representations be? And how could a Bayesian mechanism compute them?

In general, a *social state* (or situation) is a set of social variables in a process that generates sensory input. Variables are *social* when they concern features of agents' interactions. Social states are highly structured, in that the variables constituting a social state can be correlated in complicated ways. The most important of social feature is the hidden (mental) state of the other agents with whom we interact. The value of agents' hidden state both affects and is affected by the social contexts where the agents interact. Social *contexts*, recall, are sets of slowly and discretely changing parameters. These parameters comprise both slower changing variables in the internal state of agents and external variables such as features of the physical configuration of the external environment. Examples of these features are the physical arrangements of buildings and of their internal spaces. Churches, universities, cinemas, houses, parks are all examples of social contexts.

The hidden state of an agent is the most important social feature because it determines how that agent will interact with us, and how that agent will react to new sensory inputs. If we knew other agents' state, then we would have a model of their behaviour. A model of their behaviour would allow us to predict their reactions to inputs that we or the environment provide to their sensory systems. When other agents also have a model of our behaviour, we have a means to adjust our behaviour to each other by predicting each other's reactions to new inputs (Wolpert et al. 2003).

However, we don't have direct access to other agents' state. Our cognitive systems need to infer it by relying on information about the social context and about other social variables like facial expression, hand gestures, posture, physical appearance, dress, speech, tone of voice, and so on. Relying on this type of information is necessary for our computationally bounded cognitive system even if we had some direct access to other agents' internal state. Other agents' internal state, in fact, partly depends on their prior expectations about *our* state. During social interaction, their behaviour is both affecting and affected by our state. This would lead to an infinite hierarchy of priors in a computationally-unbounded agent. We are trying to infer another agent's state who is trying to infer our state: What I expect another agent's state is; what the other agent expects I expect about her state; what I expect another agent expects me to expect about her state, and so on. If we tried to infer other agents' states by using only information about mutual expectations about each other's state, then the infinity of priors about priors would make the computation of the state of the other agent unfeasible.

The approaches to this complexity can be twofold. On the one hand, our cognitive system can be thought of implementing finite rather than infinite prior hierarchies. There is evidence on strategic thinking in economic games suggesting that in fact people's hierarchy of priors about other agents' state comprises on average 1.5 levels (Camerer et al. 2004). On the other hand, the Bayesian system can constrain inference about other agents' state by relying more heavily on external social cues. All these cues need be extracted from many modalities, integrated, and combined with our prior expectations about the other agent's state. Relying more heavily on this external information can spare the Bayesian system to execute an infinite number of iterations on a hierarchy of prior expectations. After we acquire familiarity with the structure of external social cues and with the way they correlate to other agents' reactions to a given input, we need not rely on any prior about other agents' prior. The external cues would tell it all. By forming social representations from extensive interaction with certain types of external cues, we can arrive to act *as though* we knew the hidden state of other agents. Other people's reactions to a certain action would be predicted by the representation extracted from the cues present in the environment.

If this is so, then, in general, shared expectations in the form of mutual priors may not be constitutive of norm compliance. People's preference to comply with norms would not be dependent on having the right kind of mutual expectations. It would rather be dependent on the right "reading" of the cues present in situations of social interaction. This reading would in turn depend on one's acquaintance with those cues and ultimately on one's learning trajectory in the social world. Evidence about autistic people's behaviour in economic games seems to support this suggestion.

Autistic people have an impairment in their capacity to "mentalize": in their capacity to reason strategically about what other people think, feel or could do given their beliefs (Baron-Cohen 2000). Nonetheless, even though autistic people have difficulties in figuring out what other people expect, at least some central aspects of their moral knowledge and capacity to comply with norms appear to be spared in many circumstances (Blair 1996; Kennett 2002; Leslie et al. 2006, McGeer 2008). Just to give an example: Sally and Hill (2006) compared the behaviour of healthy children and adults with patients diagnosed with autistic spectrum disorder of the same age playing economic games. Games like the Ultimatum Game, where people are asked to offer or to accept/refuse a share of a certain amount of money, can be used to measure to what extent people comply with norms of fairness. Sally and Hill found that in comparison to healthy children autistic children offered significantly less in the ultimatum game, with nearly half of them offering zero or a share of one out of ten. Autistic adults, instead, showed a pattern of choices similar to that of healthy subjects.

This suggests that through extensive experience with repeated social interactions autistic subjects can build up social representations responsive to external contextual cues. The presence of certain external cues is often sufficient to activate such social representations, which enable autistic patients to implement the type of behaviour "called for" by the situation they are facing. In many situations, thus, autistics can comply with social norms, even though their capacity to mentalize with other people is impaired. An intact capacity to reason about other people's expectations facilitates our acquisition of knowledge of social norms and of social situations that call for certain behaviour. Such a capacity, however, may in general be unnecessary to enable social norm compliance. Therefore, norm compliance might in fact not depend constitutively on shared expectations. A statistical understanding of external social cues can be sufficient for people to be able to comply with social norms. After this detour, let's go back to the neurocomputational account of norm compliance I am putting forward.

According to the model I am describing, the task of computing social representations from sensory input can be mapped onto a hierarchical Bayesian model, where the lowest level represents basic physical features like displacement, acceleration, mass, orientation, and wavelength that are combined into increasingly complex representations, up to higher levels that represent social states. When the value of the prior on state Y depends on other parameters Z at higher levels, given perceptual input S_x, the resulting posterior probability is:

[4] $\mathit{Prob}\ (\mathrm{Y}, \mathrm{Z} \mid \mathrm{S}_x) \propto \mathit{Prob}\ (\mathrm{S}_x \mid \mathrm{Y})\ \mathit{Prob}\ (\mathrm{Y} \mid \mathrm{Z})\ \mathit{Prob}\ (\mathrm{Z})$

This is the simplest example of a hierarchical Bayesian model. Figure 2 illustrates a three-level hierarchical Bayesian model (modified from Shi and Griffiths 2009, Figure 2). In the example, the function that our cognitive system would have to compute is the posterior probability function *Prob* [Z | S_x] of a high-level hidden state Z given sensory input S_x. In order to carry out this computation, the system would have to reverse a *generative* (or forward) model which describes the causal process that gives rise to data assigning a probability distribution to each step in the process. Given the generative model used by the cognitive system to determine how sensory inputs are generated, the system can infer the hidden state dependent on the sensory data by reversing the generative model.

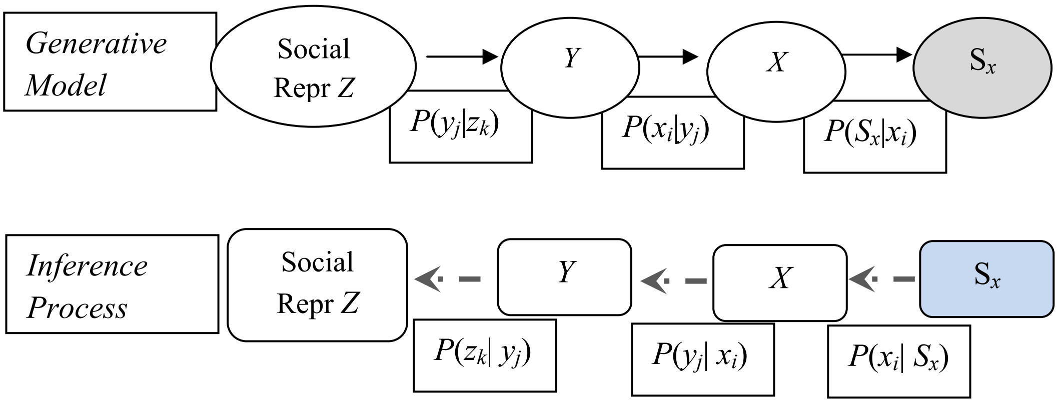


Figure 2. A Hierarchical Bayesian Model.
The generative model describes the causal process by which each variable is generated (in ovals). The inference process reverses this process (in boxes). S_x is the sensory input to the nervous system. *X*, *Y*, *Z* are neural representations at increasing level of abstraction, with *X* being the representation of some simple physical quantity like wavelength, *Y* the representation of some more abstract state like the identity of a person, and *Z* the representation of some social state like "diner in the United States."

Now, if representations of basic physical features are encoded by spikes of single neurons, more complex and abstract representations are encoded up the hierarchy by larger populations of neurons. Thus, as suggested by Eliasmith (2003, p. 503), we can build a "'representational hierarchy' that permits us to move further and further away from the neural-level description, while remaining responsible to it." Lower-level representations would systematically depend on neural transformations taking place at low-levels in the hierarchy which are directly sensitive to raw sensory inputs. Lower-level representations would be combined in a Bayesian fashion to compute more and more abstract representations at higher level. The feedback, in the form of a sensory prediction-error carried by forward connections in these hierarchical Bayesian model, would provide a means to incorporate statistical dependencies between representations at different levels of abstractions (e.g. "If the person is a waitress and I am in a diner in the Unites States, then she is likely to get angry if I don't leave a tip"). Ultimately, the dependencies between the representations and their weights in the Hierarchical Bayesian model will vary in function of one's personal learning trajectory. By interacting with waitresses in diners, for example, we shall learn to weigh certain cues more than others to update our model of waitresses' state in that type of context. While we interact with other agents, our nervous system is constantly reorganizing so that the models of the social environment it encodes get updated, and can thus serve us as maps we can use to smoothly navigate the social world.

I conclude this section by acknowledging the speculative nature of my proposal. A Bayesian mechanism of sensory perception might be extended to account for the computation of social representations. But understanding how exactly social representations are learned, encoded and updated through neural activity is enormously difficult. The neural bases of Bayesian computations have only recently started to be studied for relatively simple problems of visual perception. How exactly the brain might perform Bayesian inference and represent uncertainty in these cases is poorly understood. The problem of understanding how exactly the brain might represent *social* states involves greater challenges.

As noted by Wolpert et al. (2003, p. 596), the degrees of freedom in the state space of another agent are enormous. The fact that nervous systems are similar across people might constrain the dimension of such state space. For we might bootstrap any learning of other people's internal models by using information about the mappings between our actions and our own internal states. Yet it is often incorrect "to assign the same set of internal states to action mappings to everyone" (Ibid., p. 601). Learning the internal model of another person remains a daunting computational task. In general, we can more easily learn a model of a system by identifying its range of responses to a large range of different inputs we provide to it. But in the case of people this is typically not feasible. “[Y]ou cannot give an arbitrary battery of inputs to another person for system identification purposes, as […] another person has the option to withdraw communication once you have provided a ‘bad’ input” (Ibid.).

4. Social Norm Compliance and Reinforcement Learning

Granted that our cognitive system computes social representations in a Bayesian fashion, we need to explain how we *use* these representations to determine future movements or internal changes so as to engage in social norm compliance behaviour. We need to explain how our cognitive system tackles challenge (ii): To consume these representations to determine future movements or internal changes in the presence of, and interaction with other agents.

The second piece of neurocomputational machinery that would explain how social representations are transformed to enable us to engage in social norm compliance is the RL account of cortico-basal ganglia circuit. I already touched upon RL when I described TD-Gammon and in the Introduction. Now, I firstly describe in some detail the RL approach to cortico-basal ganglia activity. Then I explain how social reward prediction-error minimization meets challenge (ii).

RL offers models of optimal and approximately-optimal learning and decision-making in the face of uncertainty and rewards. The type of problem that RL models address can be defined by five ingredients (S, A, T, R, γ):

- States: S is the set of states which represents all possible configurations of the environment or of a system.
- Actions: A is the set of actions the agent can execute in the environment or in the system. Actions can influence the next state of the environment and have different costs and payoffs.
- State transition function: T: S x $A \rightarrow [0, 1]$ is the transition function. It specifies the likelihood of transitions from one state to the next in the environment. Given the current state s and an action a executed by the agent, $T(s, a, s')$ specifies the probability $Prob\ (s' \mid s, a)$ of moving to state s'. Note that the definition of T is typically based on the Markov assumption, according to which the transition probabilities only depend on the current state and action.
- Reward function: R: S x $A \rightarrow \mathbb{R}$ is the reward function. It specifies the reward r obtained by the agent for executing a certain action in the current state. It models the immediate costs (or punishment) and payoffs (positive reward) incurred by performing different actions in the environment.
- Discount factor: $\gamma \in [0, 1)$ is a discount rate which allows a tradeoff between short-term and long-term rewards. It specifies how much the agent cares about obtaining a given reward now rather than later in the future.

The *goal* of an agent behaving in the environment defined by S is to learn a policy function π which specifies a probability distribution over all available actions at each state such that the agent will maximize overall rewards. Goals, in general, can be conceived of as maximization of the integrated rewards obtained over many interactions within the environment. Given a change in the reward value of the choices of an agent in a state, the goal of the agent changes as well. The agent has a certain set of actions available in each state of the environment s_t. Actions give rise to a reinforcement outcome, or reward r_t, and cause a stochastic transition from state s_t to a new state s_{t+1}. Agents select actions so as to reach their goal, and maximize rewards over time.

A major problem for RL algorithms is how to balance optimally the "exploration" of the environment, to gather knowledge, and the "exploitation" of current knowledge to achieve a given goal. This problem is called "exploration-exploitation" problem. To learn about the possible outcomes of particular actions in different states, the agent must try "exploratory" actions, it has not taken yet, to expand its knowledge-base. However, the agent must also "exploit" its current knowledge to makes choices leading towards its goal. Too much exploration could lead the agent to waste time trying to have a complete knowledge of the environment instead of accomplishing its task sooner with its current knowledge. Too little exploration could lead the agent to implement an inefficient policy.

There are two main families of algorithms capable of solving the RL-problem: model-based and model-free algorithms. They differ in how they draw on experience to estimate quantities relevant to make choices and how they transform these quantities to reach a decision. *Model-based* algorithms draw on experience to build a model of the state transition and reward structure of the environment. They make choices by searching this model to find the most valuable action. Searching the model is time-consuming and computationally costly though it usually leads to accurate choices.

Model-free algorithms draw on experience to learn action values directly, without building and searching any model. Model-free algorithms don't involve much computational cost, as they need not build or search a full map of state transitions and reward structure of the environment to learn and select actions. What drives learning and action selection in model-free algorithms is a reward prediction-error of the type we already encountered in the case of TD-Gammon. This signal allows the agent to learn the value of each state $V(s)$ or the value of each state-action pair $Q\ (s,\ a)$ from trial-and-error sampling and select the action with the current highest value. Yet much training is needed in order for model-free algorithms to learn and act upon accurate value estimates courtesy of reward prediction-error minimization.

Since model-based algorithms draw upon explicit representations of state transitions and reward structure of the environment to select actions, they allow action selection to be immediately sensitive to changes in the transition contingencies and in the reward structure of the outcomes of actions. For instance, in model-based RL the tendency to select actions leading to outcomes whose reward-values have decreased is immediately diminished. Model-free algorithms, whose predictions of value only change through reward prediction-errors and slow trial-and-error experience with the environment, are instead insensitive to changes in circumstances. They cannot adapt immediately to changes in contingency and outcome reward-value.

One way to distinguish between model-based and model-free RL is in terms of goal-directed and habitual behaviour (Dayan 2009). Model-based RL underlies *goal-directed* behaviour, whereas model-free RL underlies *habitual* behaviour. These two types of behaviour have received a neat operationalization within research in animal conditioning (Dickinson 1985; Dickinson and Balleine 2002). Goal-directed behaviour "is defined as that is performed because: (a) the subject has appropriate reason to believe it will achieve a particular goal, such as an outcome; and (b) the subject has a reason to seek that outcome" (Dayan 2009, p. 213). A hungry agent pressing a button in a vending machine to obtain food is an instance of goal-directed behaviour. Here the agent has appropriate reason to believe she will get food by selecting a certain action because, say, she has experience of the contingency between that action in that situation and a certain outcome. Goal-directed behaviour is flexible since the propensity of the agent to select a goal-directed action is sensitive to manipulation of either (a) or (b). Action selection, that is, is sensitive to (a') changes in the contingency between action and outcome (for example, food is available also in the absence of the button press); and to (b') changes in the desirability of the outcome (for example, when food is poisoned, or the agent is satiated). If behaviour is not affected by these manipulations, then it is habitual. Habitual behaviour is performed repeatedly, on cue, not because of a current or future goal. It is performed because of a previous goal and the antecedent trajectory of actions that were selected to achieve that goal. Habitual behaviour occurs despite of outcome devaluation, that is, despite of the fact that the desirability of a certain outcome is reduced so that it is no longer rewarding.

RL modelling has had profound impact on neuroscience. It has helped us to understand the possible computational function of specific neural signals and patterns of brain activity (Niv 2009). In particular, the phasic activity of dopamine neurons present many of the properties of the TD reward prediction-error, which is the engine of learning and action selection in model-free RL. In the mid '90s, in fact, it was discovered that the phasic firing of dopamine neurons in the midbrain substantia nigra pars compacta (SNc) and the ventral tegmental area (VTA) can be described as encoding a reward prediction-error (Houk et al. 1995; Schultz, Dayan and Montague 1997). Like reward prediction-errors in TD-learning, the pattern of activity of dopamine neurons displays specific properties, as illustrated in Figure 3.

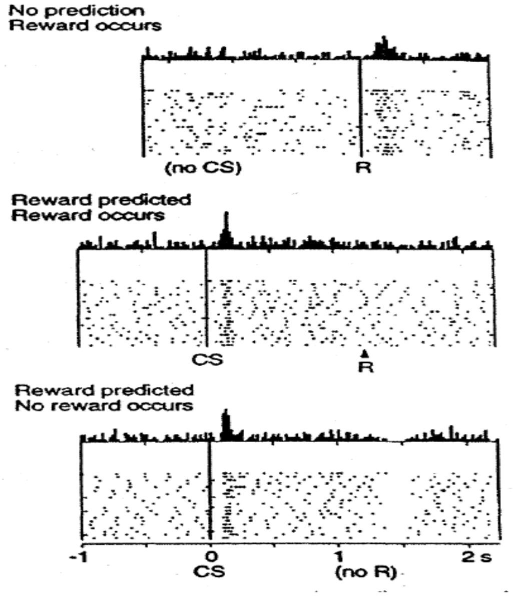


Figure 3.
Temporal Difference prediction-error and dopamine activity.
The plots show the neural activity of dopaminergic neurons of monkey during a conditioning task. The monkeys in the experiment were trained to learn that a conditioned stimulus (CS) led to a juice reward (R) a few seconds later.
On the top, is displayed a phasic burst of activity during the release of the unexpected reward (R) early in training.
On the middle, after conditioning with a cue which predicted the juice reward, the phasic burst of activity occurred at the presentation of the cue (CS) instead of the reward.
On the bottom, a dip of dopamine release when the reward was unexpectedly omitted. (Figure from Schultz et al. 1997)

"Dopamine neurons are […] excellent detectors of the 'goodness' of environmental events relative to learned predictions about those events" (Schultz et al. 1997, p. 1595). Bursts of activity in dopamine neurons occur when an agent receives an unexpected reward. In this case, dopamine activity would encode a positive reward prediction-error. After some training, once the agent has learned the association between a certain cue and the reward, bursts of dopamine activity occur when the agent is presented with the cue, as if the cue had acquired predictive value. Put differently, after a cue comes to predict a reward, it is the unexpected cue that informs you that the state in the environment is better than expected. If, at the time the reward obtains, the activity of dopaminergic neurons stays at baseline, then the predicted reward occurred as expected, and hence no prediction-error is to occur. Finally, after training in the case of a cue followed by no reward, the activity decreases thereby signaling an error in the estimate of the value of the state following the cue. This dip of dopaminergic activity might translate into a negative reward prediction-error. Since the evidence indicates that positive reward prediction-errors change dopamine firing rates more than negative reward prediction-errors (Schultz et al. 1997), it is possible that positive and negative reward prediction-errors are encoded differentially in the dopamine neurons (Bayer and Glimcher 2005); another possibility is that dopamine codes positive and negative reward prediction-errors by working together with some other system (Daw et al. 2002).

If dopamine encodes reward prediction-errors, then the input to the basal ganglia received from many diverse afferents—including the medial prefrontal cortex, the central nucleus of the amygdala, lateral hypothalamus, the serotoninergic raphe—would convey information about the outcome of a given action and the motivational significance of the current state, respectively in terms of the reward yielded by the action and the value of the current state. Dopamine neurons would transform this information to compute a reward prediction-error that is passed on to striatal target areas to facilitate prediction learning and action learning by systematically gating synaptic plasticity.

Evidence that this "cartoon picture" of the computational function of dopamine neurons might describe some aspects of the cortico-basal ganglia circuit comes both from physiology and computational neuroscience (Glimcher 2011). Plasticity in the synapses between the cortex and the striatum seems to be in fact dependent on dopamine signaling from the basal ganglia (Reynolds and Wickens

2002). Actor-Critic RL architectures, where a "Critic" supplies an "Actor" with predictions of value so that it can guide action selection, capture some known basic aspects of the basal ganglia-striatal circuits such as the phasic activity of dopamine neurons and dopamine-dependent plasticity in the striatum (see e.g. Joel et al. 2002). According to this type of architecture, dopaminergic activity in the ventral tegmental area targeting the ventral striatum (or nucleus accumbens) and frontal areas (like the orbitofrontal cortex) are used to train predictions, whereas dopaminergic signaling in the subtsantia nigra pars compacta that targets dorsal striatal areas (like the putamen) is used to learn an action selection policy.

Not all forms of learning and action selection, as modeled in RL, are dependent on dopamine however. There is both behavioural and neural evidence for a multiplicity of mechanisms of decision-making, each with computational properties suitable to different features of real-world situations, and some that do not involve dopaminergic activity. For example, Daw et al. (2005) have suggested that the central neural system might implement not only model-free RL algorithms, but also model-based algorithms. They propose that activity in the prefrontal cortex is responsible for implementing model-based strategies (thereby supporting goal-directed behaviour), whereas the dorsolateral striatum and its dopaminergic afferents would implement model-free strategies such as TD-learning (thereby supporting habitual behaviour). These two systems would represent "opposite extremes in a trade-off between the statistically efficient use of experience and computational tractability" (Daw et al. 2005, p. 1704). When the model-based and model-free strategies are in disagreement in recommending different courses of actions—Daw and colleagues argue—the criterion of arbitration used by the nervous system is based on the relative accuracy of the evaluations of the two strategies. The relative accuracy of the evaluations of the two strategies depends on such factors as the amount of training (which increases accuracy in the model-free system) and the depth of search in the model (which increases computational noise, and consequently inaccuracy in the model-base system).

4.1 RL Social Norm Compliance

I now explain how social reward prediction-error minimization meets challenge (ii), and I articulate the claim that prediction-error minimization carried out by cortico-basal ganglia circuits is a crucial component of the mechanism of norm compliance behaviour.

Prediction-error minimization of sensory input enables us to acquire social representations. But social representations, by themselves, do not motivate us to take a certain action. I suggest that the RL system bootstraps us into social behaviour and culture by transforming social representations so as to determine future movements or internal changes in the presence of, and interaction with, other people. When the RL system taps into social representations that concern the hidden state of other people, then RL system enables us to learn to comply with social norms by minimizing social reward prediction-error. An example can help me unpack these claims.

Imagine that you arrive in some foreign country. You have certain beliefs, or priors, about how situations of type *Z* look like and about how people typically behave in *Z*: you have priors concerning a social state. In particular, you have a prior over the hidden state of other people in *Z*. Yet you are uncertain about what "grammar" governs situations of type *Z* in that country, as you have a low degree of confidence about the mapping between sensory input and social representation of *Z* in that country, you are uncertain about the state transition *T* (*z*, *a*, *z'*) and you are uncertain about the reward contingencies *R*: *Z* x *A* → ℝ. If you want to interact adaptively with other people in that country in the environment *Z*, then you must learn and use the "grammar" people live by in *Z* in that country.

The first task that your cognitive system has to carry out in order to learn that "grammar" is to update your prior over the social environment *Z* in light of the information provided by the data generated by states in that environment. This task—I suggest—can be carried out courtesy of the Bayesian system described above. Let's assume that you arrive to represent *Z* as a "diner" with high confidence. By relying on this representation, you expect that people in that environment behave in specific ways since *Z* typically correlates in specific ways to the hidden states of people in *Z*. So by relying on your social representation of *Z*, you expect that the environment has a certain causal *structure*. Because you are not confident about the state transition function, and about the reward contingencies in *Z* in that new country, you have to learn them if you wish to behave co-adaptively. Assumptions about the structure of the environment can reduce the space of states and actions to a learnable subset (Gershman and Niv 2010). More generally, your expectation about the structure inherent in that environment can greatly simplify your learning and decision-making (Kemp and Tenenbaum 2009).

Now, to learn these pieces of "social grammar" your cognitive system can rely on model-based and model-free RL systems. Given limited experience with that new environment, initially you need to rely completely on an estimated model of the environment of the form:

Prob (*new state* | *state*, *action*).

This estimated model of the environment can be constrained with information about the structure associated with your social representation of *Z*. Using this model you can perform a simulation of the consequences of your actions given current state *z*: If you take action a_t from current state *z*, then it's likely that you will end up in state *z'*. You utilize experience with state transitions to update an estimated state transition function *T* (*z*, *a*, *z'*). Upon each of your choices, a *state prediction-error* is computed:

[5] $\delta_{spe} = 1 - T(z, a, z')$.

This state prediction-error is used to update the probability of the observed transition thus:

[6] $T(z, a, z') = T(z, a, z') + \eta\delta_{spe}$

where η is a parameter controlling your learning rate.

Behaviour shaped by this model-based system reflects a goal-directed process in which a particular desired outcome, like getting along or avoiding frictions with other people in *Z*, is used to flexibly determine any complex sequence of actions needed to achieve it. Action selection is carried out by searching your model of the environment: you work out the consequences of each action available to you in *z*, and select the action that is more likely to lead you towards your desired outcome. This allows action selection to be sensitive to changes in the structure of the environment and in your motivational state. If, for example, you notice that people suddenly react differently than usual given *z*, or your motivational state is abnormal, you can immediately adjust your behaviour accordingly.

Let's assume that the service in that diner was good and you have to pay your bill. You are in state z and, by relying on your prior about the structure of the environment Z you believe that most people leave a certain amount of money as a tip after their meals in a diner. You also believe that others expect you to leave a tip after your meal, as you remember that people considered you miserly when you had failed to leave a tip in a restaurant somewhere else in the past. You wish to get along with people over there, or at least to avoid frictions between you and others. You have a number of different actions available corresponding to different amounts you may leave as a tip. But money is also important to you, so the number of actions you are willing to take into account is limited. Before choosing an action, you take into consideration the likely reactions of waitresses and other diners in that state. After your choice, you observe the new state of the environment and a state prediction error is computed as in [5]. This prediction-error measures the surprise in the new state given the current estimate of the state-action-state transition probabilities. By keeping track of the specific consequences of that action as well as the causal relationship between the action and your desired outcomes, you can learn a map of the environment Z. For example, by observing others—by observing what most other diners do in Z and how waitresses react to certain actions of diners—you learn that waitresses and owners of diners over there tend to get angry when your tip is lower than a certain amount. In this case frictions between people over there ensue. Thus you can arrive to understand that people expect others to take a particular action a when they are in state z. By taking this action, it is most likely that you will satisfy your desires, thereby avoiding frictions with other people. By learning a map of Z and acting upon it courtesy of a model-based system, you have just learned to comply with a social norm.

Model-based socio-normative learning is supported by a distributed neural mechanism. “The behavioral neuroscience of such goal-directed actions suggests a key role in model-based RL (or at least in its components such as outcome evaluation) for the dorsomedial striatum (or its primate homologue, the caudate nucleus), prelimbic prefrontal cortex, the orbitofrontal cortex, the medial prefrontal cortex, and parts of the amygdala” (Dayan and Niv 2008, p. 186). Specifically, using functional magnetic resonance in humans, Gläscher et al. (2010) found a neural correlate of a state prediction-error in the intraparietal sulcus and lateral prefrontal cortex. This finding supports the existence of a unique learning signal in the brain, which, apart from guiding model-based learning and action selection, seems to drive learning of causal relationships between cues and consequences as well. The minimization of state prediction-errors in social situation *Z* is what drives your learning to comply with a social norm in *Z*.

Activity in the neural structures just singled out support “effortful” computational processes. Searching and updating your “map” of the environment is in fact computationally demanding both for working memory and for your “mentalizing competence.” You need to remember situations you encountered in the past similar to the one at hand, you need to work out what other people’s expectations may be, you have to consider many different actions and outcomes, and work out which is the best to achieve your goals. This can reduce the capacity for alternative computations and the smoothness of interaction, as the model-based system would engage valuable cognitive resources to identify which action you should implement given your state and your goals. By relying on a model-based controller, learning and complying with social norms can be effortful and time consuming.

One crucial aspect of social-decision problems is that they typically recur. So with more experience with situation *Z* in that country, you need not to rely on the model-based system. After you have regularly encountered situations of type *Z*, the sensory data generated by *Z* have led your representation of *Z* to be more and more accurate. Thus your prior about the structure of that environment can impose further constraints on the state and action space, on which your learning and decision-making systems tap. Now you can rely on a model-free system which drives learning and decision making by means of social reward prediction-errors encoded by dopamine activity. The reward is social because it is brought about by other people's reactions to your behaviour: they may openly or more subtly approve or disapprove of your behaviour. By picking up on these rewards, you acquire ways of evaluating or predicting the long-term consequences associated with executing a particular action. You need not "mentalize" with others or search any map of the environment. You can come to comply with social norms automatically, quickly and at little computational costs.

The model-free system can operate effectively with little computational demands in familiar situations. This system operates on "cached" values that store experience about the overall future worth of a particular action. Such values can be used to implement certain behavioural responses in the face of stimuli that were consistently associated to a rewarding outcome in the past. Given reliable co-variation between situational cues and certain behavioural patterns of people in *Z*, the reward values of the behavioural responses become conditioned onto the cues. Features of the environment become to encode information about the reward structure of the environment, and you can outsource behavioural control on them. The cues present in the environment signal opportunities to perform particular "rewarding" actions. In this way, as your training with social situation *Z* proceeds, goal-directed behaviour becomes habitual and cue-driven. The representation of *Z* itself can drive behaviour with no need to work out what other people expect you to do in *Z* or to keep track of state transitions underlying *Z*. Features of *Z*, that is, acquire the capacity to motivate you to directly act upon your social representation of *Z*. Norm compliance in this case becomes perceptually-based.

Note, however, that the shift from model-based to model-free control is not sequential nor instantaneous, but highly parallel and dynamic. The early phase of model-free learning processes take place while behaviour still appear to be controlled by a model-based system. Tricomi et al. (2009) provide evidence of the dynamic recruitment of both model-based and model-free controllers during human decision-making. In their imaging experiment they found that the dorsolateral striatum—which is thought to support habit-learning—increases gradually, and not suddenly, over training in a task that initially calls for model-based control. Furthermore, activation of the ventromedial prefrontal cortex—which is thought to support goal-directed behaviour by representing the value of action outcomes—was observed throughout all training sessions in their experiment in anticipation of reward outcomes. Habitual behaviour then seems to result not from repetition of a certain action *per se*, nor from a decrease in the anticipation of reward outcomes, but rather from the fact that extensive experience with a certain environment enhances the sensitivity to cues associated with a particular behavioural response. Perception of environmental cues may thus directly drive us to comply with social norms. And complying with social norms by acting upon our perception of environmental cues facilitates us to behave smoothly and adaptively in the presence of others.

In sum: Our acquisition of the grammar that governs social situations can then be driven by minimization of three types of prediction-error: a sensory prediction-error that is produced and minimized by a Bayesian system, which gives rise to social representations; a state prediction-error that is produced and minimized by a model-based RL system; and a social reward prediction-error that is produced and minimized by a model-free RL system. These two RL systems enable us to act on our social representations so that we comply with social norms. Bayesian and RL algorithms may be implemented by cortico basal-ganglia circuits, where dopamine plays a central role in both learning and acting upon certain representations.

By working in concert, such a Bayesian-RL neurocomputational system ensures that our predictions about people's behaviour become self-fulfilling prophecies. Our complying with norms is one trick we use to make these predictions come true in social environments. It ensures that our prior expectations about social sensory input are met and social uncertainty is avoided. When norm compliance becomes a habit, governed by a model-free system, social interaction becomes a fluid, flexible, context-specific, inferential response to incoming sensory input and their values. It enables co-adaptive, smooth interaction without access to hidden states of other agents in the world.

5. Bayesian-RL Neural Computing as Building Blocks of Norm Compliance. Three arguments

I conclude this chapter by laying down three arguments for why the type of neurocomputational model I put forward should be used as a framework for understanding norm compliance behaviour. First, such neurocomputational model is supported by evidence from the neuroscience of social decision-making. Second, my neurocomputational model explains the nine core features of norm compliance identified in the Introduction. Third, understanding social norm compliance within a Bayesian-RL framework has an advantage over competing, non-computational, accounts since it a) warrants us from arbitrary descriptions and predictions of phenomena, and b) fosters integration of individual findings about social norm compliance from different disciplines. These three reasons are articulated in turn.

5.1 Neural Evidence for a Bayesian-RL mechanism of norm compliance

There is a substantial body of evidence that the neural circuits of the Bayesian-RL system I described are not only involved, but they also might be essential for the acquisition of and compliance with social norms (Fehr and Camerer 2007; Lee 2008). In this section I illustrate this claim with two examples.

Game theory is the most widely used formal framework for studying social interactions. Recently, games such as the Prisoner's Dilemma, the Trust Game and the Ultimatum Game have begun to be combined with technologies and methods from the cognitive neurosciences like brain imaging. From these studies, the theme common to social decision-making is the basal ganglia-based circuit. This circuit has widespread connections with limbic and sensorimotor mechanism. As pointed out by

Fehr and Camerer (2007, p.422), there is an apparent overlap between the areas like the orbitofrontal cortex and other prefrontal regions, and the dorsal and ventral striatum activated in tasks where social norms shape subjects' behaviour, and activations observed in studies of reinforcement and habit learning.

Spitzer et al. (2007), for example, asked how the brain may process the threat of punishment when we decide whether or not to comply with a social norm. To answer this question they used fMRI while their subjects played a trust game where norm violation could be punished. In this game, player A (in the fMRI scanner) was given a sum of money which he could distribute between himself and player B who was anonymous. In the control condition, player B was a passive recipient of A's offer. In the "punishment threat condition" player B could punish A after A's offer was revealed. Player B had a monetary endowment which he could spend to reduce A's payoff. The threat of punishment made people act more fairly. In the "punishment threat condition" people split the money close to equally. When player B had no recourse, the people who were given the money acted differently and gave away, on average, less than 10 percent of the money.

Individuals' increase in norm compliance under the "punishment threat condition" correlated with activations in the lateral orbitofrontal cortex, right dorsolateral prefrontal cortex and caudate. Lateral orbitofrontal cortex activity was also found to be correlated with "Machiavellian personality traits" which were previously measured with a questionnaire. This questionnaire aimed to measure each subject's combination of selfishness and opportunism. Notice that subjects with high Machiavellism scores gave less money in the control condition and were best at avoiding punishment in the "punishment threat condition". The orbitofrontal cortex, then, was most activated in the more selfish and opportunistic subjects. These results suggest that the role of the orbitofrontal cortex is to enable people to detect and evaluate social cues such as the threat of a punishment.

This type of result can receive a natural interpretation by appealing to computations of reward prediction-errors. Because BOLD signals may not directly reflect firing activity in a certain region, it is more appropriate to consider imaging results as reflecting the information that that region is receiving and processing, rather than the information transmitted to downstream targets (Niv and Schoenbamum 2008). So BOLD signals found in striatal and prefrontal cortical areas, which are primary target of dopamine neurons, may encode the information carried by prediction errors computed in the basal ganglia.

When target areas of reward prediction-errors are damaged from early age, our cognitive system might have difficulties in acquiring and acting upon social representations. Hence social behaviour and our capacity to comply with norms can be compromised. Anderson et al. (1999) studied two adult subjects whose ventral, medial and dorsal regions in the prefrontal cortex were damaged before sixteen months of age. These two patients exhibited an inability to interact adequately with other people *and* they could not retrieve explicit socio-normative knowledge. Because of dysfunction in cortical areas that might be necessary for Bayesian computation of social representation, the two subjects might never have acquired socially relevant knowledge in spite of extensive exposure to a variety of social information in their home and school environments. Treatment with social programs aimed at correcting their inappropriate behaviour during adolescence was unsuccessful. So their incapacity to comply with norms might have depended on incapacity to perceive and respond adequately to cues present in a given situation, and this incapacity might ultimately depend on an inability to compute social representations.

5.2 Nine features of norm compliance explained

The Introduction identified nine seemingly core features of social norms or social norm compliance:

I. Norm compliance depends constitutively on shared, mutual expectations.

II. Norm compliance is intimately related to punishments and rewards.

III. Norm compliance is conditional on having the right kind of representations.

IV. Norm compliance does not depend on a supply of invariant general principles.

V. Social norms set the boundaries of "appropriate" behaviour.

VI. People are subject to many sources of motivations.

VII. Social norms have special motivational grip.

VIII. Complying with norms is thoughtless.

IX. Socialization is necessary for the development of norm compliance.

A Bayesian RL model of norm compliance behaviour would explain, or explain away, all these features. Let me briefly consider each feature in turn.

5.2.1 Why does norm compliance *seem* to depend constitutively on shared, mutual expectations?

A person's expectations can be described as the hidden state of that person. People's hidden states are the most important social representations, as they directly determine how those people will interact with us and how they will react to new sensory inputs.

If in a given social situation we have reliable expectations about other people's expectations, and other people have reliable expectations about our expectations, and these mutual expectations are common knowledge, this information can be used to behave co-adaptively. Therefore, norm compliance *seems* to be constitutively dependant on mutual, shared expectations.

However, as I argued in section 4.3, norm compliance is probably *not* constitutively dependent on mutual, shared expectations. We need not directly infer other people's hidden state—which would be a computationally daunting task. The same type of information can be outsourced on external cues. By computing social representations from these cues, we can behave *as though* we acted upon each other's expectations.

5.2.2 Why is norm compliance intimately related to punishments and rewards?

Because the capacity to act upon social representations depends on the workings of RL systems and model-free of RL systems. Such systems bootstrap us into a world of culture courtesy of social reward-prediction errors. At the level of RL systems, rewards and punishments are units of information which may not be identical to positive or negative feelings. These units of information colour by association otherwise neutral states in our social environment as states to be approached or avoided, states we care or care not about. The stamping-in of reward values to states in our environments is driven by prediction-errors and is ultimately a function of the goal of adapting one's behaviour to other people's behaviour.

5.2.3 Why is norm compliance conditional on having the right kind of representations?

Because norm compliance depends on perceptual skills. We acquire social representations courtesy of a Bayesian system. That we have representations of the "right" kind means that the perception of a given social situation yielded by the Bayesian system is such that co-adaptive, fluid behaviour is likely to ensue if we act upon this perception. When we misperceive a given situation, the probability of social misbehaviour arises. Failing to correctly perceive a social situation is likely to cause a failure in norm compliance thereby engendering frictions with other people. Ultimately, in a given situation, a social representation is "accurate," or "right" in so far as the information it encodes reliably correlates with other people's hidden states in that situation.

5.2.4 Why does social norm compliance not depend on a supply of invariant general principles?

Because our social world is a dynamic, complex system. Features which function as drives of social norm compliance in one situation at a time need not function as drives of norm compliance at all in another situation or at another time. The way these features function at a given place and time can be described by means of a general principle. But this does not mean that people are always motivated to comply with a social norm at a given time and place because of that general principle. Two classes of rules by which we navigate this world consist in Bayesian inference and RL algorithms. These rules enable us to perceive certain social patterns at a given time and constrain the ways our social world changes over time because of the way we perceive each other and act in the social world. Note that in general changes of our social world, or cultural evolution, can occur as effect of many factors: we may experiment with new behaviour; we may consciously start to imitate somebody else; we may instruct our children in certain ways; there may be random fluctuations of people's beliefs and expectations; by migrating and meeting other people certain beliefs and expectations may be introduce or eliminated in a given place and time.

5.2.5 Why do social norms set the boundaries of "appropriate" behaviour?

The basal ganglia-based reward system is a device for leading agents to approach certain things rather than others. It estimates the reward value of acting upon one stimulus-representation rather than another, and thus it prepares a certain motor response. The reward value of a given representation is in function of the goal of co-adaptive behaviour. Pursuing rewarding social states is one way we come to comply with norms and thereby we can behave co-adaptively and fluidly. So by pursuing rewarding social states, we behave "appropriately." When we fail to pursue social states with high reward value, it is likely that we fail to comply with social norms. Thus we may behave "inappropriately." Social norms and appropriate (inappropriate) behaviour can be described in function of the reward-value of social states in a given environment.

5.2.6 Why are people subjects to many sources of motivations?

Because our learning and decision-making are driven by multiple systems with different computational properties, and different types of circumstances favour different systems. Combining different systems can thus be advantageous given the characteristics of the diverse environments where we behave. I have described two systems for learning and decision-making: model-based and model-free RL. The circumstances that favour model-based system are typically those in which we do not have sufficient experience such as when we face a new social situation in a foreign country. Because social-decision problems recur, when we are familiar with a certain situation, behaviour tends to be driven by model-free evaluations.

5.2.7 Why do social norms have special motivational grip?

Our social nature compels us to pursue co-adaptive, frictionless, fluid behaviour with other people. This goal has high-value to us. Because complying with norms is one prominent way we have devised to pursue this, social norms have special motivational grip. Furthermore, norm compliance serves best the goal of co-adaptive, frictionless social behaviour when it becomes a kind of habit. Habits have special motivational grip in that they are resistant to devaluation. So when norm compliance becomes a habit, it acquires extra-motivational grip in that it is resistant to devaluation.

5.2.8 Why is norm compliance thoughtless?

Because norm compliance is paradigmatically governed by a model-free, habitual system that involves little computational cost in terms of neural resources and time. When norm compliance becomes a cue-triggered, habitual response, action-selection is computationally cheap and automatic. In this sense norm compliance becomes thoughtless: we comply without thinking about it.

6.2.9 Why is socialization necessary for the development of norm compliance?

Because social information is necessary for the Bayesian system to yield social representations and for the RL systems to implement certain actions so that we can comply with social norms. Malfunctioning of the Bayesian system or development in situations of social deprivation can engender incapacity to acquire complex social representations on which we have to rely to comply with social norms. Malfunctioning of the RL systems or being raised with a lack of caring relationships can engender incapacity to being sensitive to the reward values of the social states of a certain environment. Thus, although we may still have a data-base of social knowledge, we may fail to act upon it, as we may fail to attach any reward value to social representations thereby becoming motivationally insensitive to social representations.

5.3 Virtues of a Neurocomputational Perspective

There are already empirically informed models of social norms and social norm compliance. Both Bicchieri (2006) and Sripada and Stich (2006)—just to name two works on social norms carried out by philosophers—develop frameworks for the study of norms and norm compliance by relying on findings from social psychology, experimental economics, cognitive neuroscience and anthropology.

Sripada and Stich's (2006) model is a "boxological" model of the mechanism underlying the acquisition and implementation of norms. Their model describes a set of functionally individuated components (black boxes) underlying such mechanisms, the processes they go through, and their organization. Bicchieri (2006) proposes a model of norm compliance as the product of expected utility maximization by socialized, boundedly rational agents. Some concerns can be raised about both approaches. For example, it is controversial whether a boxological approach can in fact yield genuine explanations, as the boxes it postulates are not identified with concrete structures and internal states of a system; moreover it is not obvious to what extent "rational reconstructions" such as Bicchieri's accurately model the psychology of individuals, as it is unclear to what extent the parameters posited in their utility functions pick out features of people's psychological make-up.

These models, nonetheless, remain valuable tools for further research on norms. For at least they offer us with frameworks that can be used to understand known phenomena about social norms and to test new hypotheses about norm compliance behaviour. What would a neurocomputational model of norm compliance bring to the table?

There are several advantages of expressing a model of norm compliance in equations which aim to provide approximate descriptions of some of the features of its neurobiological mechanism. I focus on two virtues of neurocomputational models. First, the inferences we draw about the target system represented by the model are typically *non-arbitrary*. Second, neurocomputational models foster *integration* of disparate phenomena studied in different disciplines.

If the inferences drawn from a model about its target system are arbitrary, then that model cannot reliably be used to describe, predict or explain certain phenomena concerning the target system. Self-consistency is a warrant against arbitrariness. Inconsistent models cannot be (approximately) true descriptions of some features of a mechanism. So if neurocomputational models are used to extract non-arbitrary descriptions of some features of their target systems, they must be self-consistent. Neurocomputational models are expressed in equations which require a precise, quantitative, self-consistent formulation. So neurocomputational models can be used to extract non-arbitrary descriptions of some features of a mechanism (Abbott 2008).

It might be difficult to extract non-arbitrary, quantitative predictions about the outcomes of a mechanism in different situations from boxological models or "rational reconstructions" of norm compliance. In comparison to these two approaches, neurocomputational models are more explicit and precise in their commitments. The mathematical formulation of some ideas about the functions carried out by neural circuits underlying norm compliance behaviour allows us to completely work out the consequences of the model. It allows us to formulate quantitative predictions that can shed light on the neural or the informational constraints of the mechanism of norm compliance that the model represents. By incorporating knowledge of such constraints and of mechanistic details, neurocomputational models can make informative predictions that generalize across situations. This is one way neurocomputational models of norm compliance can become genuinely explanatory in that they can come to describe the relationship between neural responses and the stimuli that evoke them on the basis of known physiological features of our cognitive systems. These types of descriptions correspond to mechanistic explanations which allow us to identify which organized structures and processes are essential for norm compliance behaviour.

With their search for basic principles that could guide us through the complexity of the neural circuits and cognitive functions of social norm compliance, neurocomputational models foster *integration* of disparate phenomena and theories in different fields. Linking various phenomena studied in different disciplines in the behavioural sciences can constitute a fruitful discovery heuristic and a force towards inter-theoretic coherence.

The basic principles used to account for a given set of phenomena can be used to understand a distinct set of phenomena by suggesting concepts to make sense of those phenomena, and new hypotheses that could be tested empirically. If the same basic principles account for two distinct sets of phenomena studied in different fields, then those two sets might not be disjoint and information about one set could be used to inform, constrain, reconfigure and displace existing taxonomies used in both fields. A neurocomputational model of norm compliance could rely on the same basic principles used to account for solutions in cognitive domains such as perception, motor control and learning (Wolpert, Doya and Kawato 2003; Behrens et al. 2009). By relying on basic computational principles such as Bayesian inference, scientists can use related research in one field to stimulate discovery at another. The use of the same computational principles can generate research that leads to the development of hypotheses about the connections between particular models employed to account for distinct phenomena like motor control and social interaction. Uncovering and developing connections between different phenomena studied in different fields amount to carrying forward a co-evolutionary research ideology, whereby research in one field can draw on concepts, empirical findings, and methodological tools from another field (Churchland 1986, Ch. 5).

Findings relevant to norm compliance from behavioural economics, social psychology, social neuroscience, biology, anthropology and artificial intelligence could be understood within one explanatory framework grounded in “basic principles” like efficient coding, Bayesian inference, adaptive optimal control, and generative models (Abbott 2008). Such an explanatory framework could serve as a bridge between models and micro-theories of the various disciplines. Cross-disciplinary links will force overlapping theories and models to cohere with each other (on the value of unification in the behavioural sciences see e.g. Gintis 2007). As explained by Pat Churchland, “[t]he unity of science is advocated as a working hypothesis not for to sake of puritanical neatness or ideological hegemony or hold positivistic tub thumping, but because theoretical coherence is the ‘principal criterion of belief-worthiness for epistemic units of all sizes from sentences on up’ (Paul M. Churchland 1980). Once a theory is exempt from having to cohere with the rest of science its confirmation ledger is suspect and its credibility plummets. To excuse a theory as hors de combat is to do it no favors” (P.S. Churchland 1986, p. 376).

CHAPTER 2.

A Plea for Neural Representations

NICE GUY EDDIE
Okay, everybody cough up green for
the little lady.
Everybody whips out a buck, and throws it on the table.
Everybody, that is, except Mr. Pink.
NICE GUY EDDIE
C'mon, throw in a buck.
MR. PINK
Uh-uh. I don't tip.
NICE GUY EDDIE
Whaddaya mean you don't tip?
MR. PINK
I don't believe in it.
NICE GUY EDDIE
You don't believe in tipping?

Both philosophical and ordinary explanations of social norm compliance generally make fundamental reference to beliefs and preferences (or desires).[1] But how should we understand the claim that people comply with a norm because they possess the right kinds of beliefs and preferences? The previous Chapter articulated a subpersonal explanatory framework and claimed that neural representations are an essential ingredient of explanations of norm compliance. Does this mean that we should understand beliefs and preferences in terms of neural representations? This chapter defends two claims:

1) The explanation of paradigmatic cases of norm compliance behaviour requires the appeal to representations. Hence, if computation requires representation, we would have independent support for explaining norm compliance from a

[1] Decision theorists tend to talk of ‘preferences’ instead of ‘desires.’ In what follows I use ‘preference’ and ‘desire’ interchangeably, as my argument does not hinge on any distinction between them. Chapter 6 will further elaborate on the notions of preference and desire.

neurocomputational perspective since both explanation of norm compliance and neurocomputational explanations would require representations.

2) The appeal to belief and preference (or desire) in explanations of norm compliance is more fruitfully understood as an appeal to neural representations rather than to behavioural dispositions. In this sense, people comply with norms because they possess the right kinds of neural representations.

It is important to clarify at the outset the dialectic underlying this chapter. I do *not presuppose* the existence of beliefs and preferences (or desires) as folk-psychological states. Rather, I explicate how the notions of belief and preference are employed in computational neuroscience in terms of neural representations, and examine their explanatory purchase. My argument is as follows: If beliefs and preferences are fruitfully understood in terms of neural representations, and positing neural representations gives non-trivial explanatory purchase with respect to norm compliance, then there is reason to appeal to neural representations in explanations of norm compliance. The argument presupposes that "explanatory relationships are relationships that are potentially exploitable for purposes of manipulation and control" (Woodward 2003, v). Accordingly, I presuppose that the adequacy of the explanatory relationship between belief-preference and norm compliance can be assessed in terms of the type of control and manipulations of norm compliance that such a relationship can facilitate.

There are five sections in this chapter. Section 1 sets the stage by rehearsing Cristina Bicchieri's (2006) theory of norms. Bicchieri extends the seminal contributions of David Lewis, Philip Pettit and Bob Sugden in analyzing social norms by using the tools of belief-preference rational choice theory. In order to explain norm compliance, Bicchieri conceives of beliefs and preferences as *behavioural dispositions*. Her account is useful to introduce different views about what it takes to have a belief or a preference. In particular, her account is useful to highlight the explanatory relationships between norm compliance behaviour and different ways to conceive of belief and preference.

Section 2 starts to put into focus the second claim defended in the chapter: it describes a case-study from computational neuroscience, and explicates the notions of belief and preference as neural representations typically assumed in computational neuroscience.

Section 3 puts forward a first argument for a version of representationalism. This argument relies on Clark and Toribio's (1994) notion of "representation-hungry" problem domain, and aims to show that explanation of paradigmatic cases of norm compliance behaviour requires the appeal to representations.

Section 4 tackles the objection that norm compliance does not consist in behaviour in representation-hungry domains by engaging with some aspects of Hubert Dreyfus's anti-representationalism.

Section 5 articulates an independent argument for neural representationalism. *If* to have a belief or a desire is to have some neural representations rather than certain behavioural dispositions, then belief-desire explanations of norm compliance are especially apt to facilitate control, manipulation or prediction. Hence there is reason to prefer representationalism over certain versions of dispositionalism as an account of the beliefs and preferences featuring in explanations of norm compliance.

A few *caveats* before getting started: my target is neither Bicchieri's account of social norms nor belief-desire "folk" psychology. For, on the one hand, Bicchieri's account remains silent about representations. On the other hand, the issue here does not concern the folk concepts of 'belief' and 'desire;' it concerns actual mental states and their explanatory relationship with norm compliance. My target is not any dispositionalist account of belief/desire either. I do not claim that any type of dispositionalist view of belief and desire is inconsistent with (neural) representationalism. I construe dispositionalism so as to better highlight the explanatory fruits of neural representationalism as a way to understand belief and desire. My general target is *any anti*-representationalist view according to which cognition and behaviour need *not*, and sometimes are *not*, to be explained in terms of representational structures and transformations over such structures.

1. Belief, Preference and Norm Compliance

As already noticed, Cristina Bicchieri offers a "constructivist" account of social norms, "one that explains norms in terms of the expectations and preferences of those who follow them" (Bicchieri 2006, p. 2). The basic idea is that "the very existence of a social norm depends on a sufficient number of people believing that it exists and pertains to a given type of situation, and expecting that enough other people are following it in those kinds of situations" (Ibid.). Social norms are social, for Bicchieri, because we prefer to comply with them *only if* we believe that most members of our society will do the same and we believe that most members of our society expects us to follow that norm.

What is important for my purposes is her claim that "the belief/desire model of choice […] does not commit us to avow that we always engage in conscious deliberation to decide whether to follow a norm. We may follow a norm automatically and thoughtlessly and yet be able to explain our actions in terms of beliefs and desires" (Ibid., p. 3). Bicchieri's argument for this claim is the following.

P1. Norm compliance behaviour does not generally involve deliberation.

P2. Deliberation involves "beliefs and desires of which we are *aware*" (p. 6).

C1. Norm compliance behaviour does not generally involve beliefs and desires of which we are aware.

C2. If beliefs and desires feature in the explanation of norm compliance behaviour, then they do not generally feature as conscious mental states (i.e. states of which we are aware).

P3. A *dispositionalist* account of beliefs and desires does not conceive of beliefs and desires as conscious mental states.

C3. If beliefs and desires feature in the explanation of norm compliance behaviour, then they can be conceived of as "*dispositions* to act in certain way in the appropriate circumstance" (p. 6).

Bicchieri begins by pointing out that most of the time we follow norms thoughtlessly, by relying on heuristics of which we are unaware. Heuristics are rules of thumb that can solve cognitive problems in little time and with little information. Heuristics, for Bicchieri, can underlie norm compliance by activating default rules cued by contextual stimuli. From this perspective, "norm compliance is an automatic response to situational cues that focus our attention on a particular norm, rather than a conscious decision to give priority to normative considerations" (Ibid., p. 5).

Bicchieri then contrasts the heuristic route to behaviour with deliberation. "Deliberation—Bicchieri writes—is the process of consciously choosing what we most desire according to our beliefs" (Ibid., p. 6). If beliefs and desires are conscious mental states, then—Bicchieri goes on—they cannot play a role in the heuristic route to norm compliance, and they cannot generally play an explanatory role in norm compliance, as norm compliance is generally automatic, effortless and unconscious. Beliefs and desires, however, need not be conscious states. Therefore they can feature in our explanation of norm compliance even when behaviour is guided by heuristics. To motivate her position, Bicchieri embraces a dispositional account of belief and desire according to which beliefs and desires are *dispositions to act* in certain ways under appropriate circumstances. She characterizes what is to believe and to prefer thus: "to say that someone has a belief or preference implies that we expect such motives to manifest themselves in the relevant circumstances" (Ibid.).

Dispositionalism allows us to rely on preferences and beliefs for the explanation of norm compliance both when norm compliance is the outcome of deliberation and when it comes from the "heuristic route." So, the fact that beliefs and desires should often feature as unconscious mental states in the explanation of norm compliance suggests that we can conceive of them as behavioural dispositions, since a dispositionalist account of belief and desire does not commit us to see belief and desire as mental states of which we are aware. Note that Bicchieri doesn't claim that we *should* embrace dispositionalism; she doesn't claim that what is essential to believing and preferring is the disposition to act in certain ways under certain circumstances. All she claims is that this type of dispositionalism is a natural way to conceive of beliefs and desires as unconscious states.

But, if dispositionalism is not the only available option to make room for unconscious beliefs and desires in the explanation of norm compliance, then we may consider independent explanatory payoffs of alternative accounts of what it is to believe and desire something. The remainder of this section argues that a dispositional account of what is to believe is not needed to explain norm compliance.

1.1. Dispositions and Representations

For a dispositionalist what is essential to belief and preference is a certain pattern of potential and actual, verbal and nonverbal behaviour under appropriate circumstances. "For someone to believe some proposition *P* is for that person to possess one or more particular behavioral dispositions pertaining to *P*" (Schwitzgebel 2006/2010). In this sense, to say that Mr. Pink believes that *P* or desires that *Q* is on a par with saying that salt is soluble, or that your supervisor is irascible, or that glass is fragile. Dispositionalists are committed to the claim that having internal representations is *not* what is essential to possess mental states. Internal representations are only relevant to the extent that they underwrite behavioural dispositions; they do not ground explanations of behaviour in terms of belief and preferences.

Dispositionalism, in comparison to the view that believing (and preferring) is to having internal representations, seems to have a difficulty in distinguishing between those cognitions that are *explicit*, those that are *implicit* and those that are *tacit* (Haugeland 1998, Ch. 7). Appealing to representations, instead, provides a useful way to put into focus such distinctions.

One cognitive system has the *explicit* belief that *P* (or desire that *Q*) if it explicitly possesses cognitive states that carry the right sort of information tokened in it. If beliefs and desires are understood as representations, then one has the explicit belief that *P* (or desire that *Q*) if some representational structure with the right sort of content is stored in the cognitive system. For example, Mr. Pink has the explicit belief that everybody is leaving a dollar on the table at the restaurant if a representation with that content is tokened in his cognitive system in the right way.

Beliefs and desires are *implicit* if they are not actually tokened in the system, but are swiftly derivable from explicit beliefs and desires in the cognitive system. In terms of representations, the distinction between explicit and implicit belief depends on whether the right representation is tokened in the system or not. Yet, as swiftness is a matter of degree, "there will not be a sharp line between what one believes implicitly and what, though derivable from one's beliefs, one does not actually believe" even implicitly (Schwitzgebel 2006/2010). For example,[2] Mr. Pink may want to leave a big tip for the waitress and believe that big tips impress waitresses. Mr. Pink held those mental states explicitly, but he doesn't draw any logical implication—though his system could swiftly draw it. Thus, we can say that Mr. Pink also wants, implicitly, to impress the waitress.

'Tacit' is used differently by different authors (Cf. Dennett 1982; Engel 2005; Fodor 1968). Here, by 'tacit cognitions' I refer to a kind of competence built into the system and evinced from the behaviour emerging from the workings of the whole cognitive system. Tacit cognitions are neither explicitly tokened nor implied by explicit representations. For example, if people's performance in a number of perceptual tasks approximates Bayesian inference, it can be said that those people are sometimes *tacit* Bayesian observers. Any one component of their cognitive system

[2] The following parallels an example in Haugeland (1998, p. 143).

need not map onto a single component of the Bayesian model. Instead, it is the cognitive system as a whole that performs Bayesian inference.

'Tacit,' 'explicit' and 'implicit' are to be distinguished from 'conscious' and 'unconscious.' *Conscious* beliefs are those that occur when people consciously entertain them. In representational terms, when Mr. Pink is asked to leave one dollar for tip, he accesses and retrieves some of the relevant representations stored in his cognitive system. He then consciously entertains the belief that all the other guys are leaving a dollar for tip. Thus, we can become conscious of beliefs and desires of which we were previously unaware. In a different sense, some mental states or processes are unconscious just in case they *cannot be accessed*. Thus, even if Mr. Pink tried to uncover the types of algorithms implemented by his brain activity when he learns a new social norm, he wouldn't be able to have access to them. Identifying such processes would take deep, systematic investigation at both the personal and the subpersonal level.

There can be explicit beliefs that are inaccessible to consciousness. In representational terms, one has explicit beliefs that are inaccessible to consciousness if there are representations tokened in the system carrying the right sort of information, but that cannot be accessed or retrieved. Chomsky (1980), for example, argues for this possibility when he talks about the representation of a grammar in our head. In the sense employed here, these types of unconscious, inaccessible beliefs are not tacit since they are actually tokened in the system.

One last important distinction, which can be drawn in terms of representations, is between occurrent and dispositional mental states. We may say that Mr. Pink *dispositionally* believes that most people leave a tip in restaurants if he has a representation with that content stored in his head *but* that representation has currently not been retrieved for active deployment for reasoning or decision-making. When, given eliciting circumstances, that representation is accessed and retrieved for active thinking and decision-making, Mr. Pink *occurrently* believes that most people leave a tip in restaurants. It should be clear then that "one needn't adopt a dispositional approach to belief in general to regard some beliefs as dispositional in the sense here described" (Schwitzgebel 2006/2010). Bicchieri's argument seems to understand beliefs and desires as *behavioural dispositions*. But to have beliefs and desires as behavioural dispositions is distinct from having representations that are dispositional *viz*. non-occurrent. To say that most of our beliefs and desires are dispositional does not entail a dispositionalist view of what it takes to believe and desire. One can maintain that representations of some sort are essential to believe and desire and still acknowledge that most of these representations are unconscious or dispositional.

Bicchieri suggests that a dispositionalist account of belief and desire fits nicely with the heuristic route to norm compliance. But "the heuristic way to behaviour" fits nicely also with a representationalist account of belief and desire since representationalism can also account for unconscious beliefs and desires. Furthermore, unlike dispositionalism, it seems that representationalism can make good sense of the distinctions between explicit, implicit and tacit cognitions. Therefore, a dispositionalist account of belief and desire is not required to allow for the fact that we are not aware of most of our beliefs and desires; and, in comparison to representationalism, it has probably more difficulty in drawing important distinctions between different types of cognitions.

2. Beliefs and Preferences in Computational Neuroscience

Nice Guy Eddie's decision to leave a tip after his meal is an example that involves a *social preference*. Mr. Pink's decision to not leave a tip is also an example that involves a social preference. Theories of social preferences are concerned with how people make decisions when the outcomes of those decisions impact the outcomes of other people. In the last decade, a wealth of behavioural and neural data has been collected about how people make social decisions. Such data, together with the ideas of social preference and bounded rationality, are beginning to be modeled and put at work in computational neuroscience. I now describe a case study from this field in order to explicate one way to understand beliefs and preferences.

Ray et al. (2009) used a Bayesian framework to model important aspects of social decision-making. They focused on a multi-round, sequential Trust Game. In each round of a Trust Game, an agent (the investor) decides how much money out of an initial endowment to send to another agent (the trustee). This amount is multiplied by some factor—e.g. three—and then the trustee decides how much of the money received to send back to the investor. Both investor and trustee know that the game terminates after a certain number of rounds. The standard game-theoretic prediction for a single, anonymous interaction between two narrowly self-interested, rational agents is for the investor to send nothing since the investor should anticipate that the trustee will not reciprocate. Experimental results, however, are inconsistent with this prediction. The average investor sends a significant amount of the initial endowment, and most trustees reciprocate (Camerer 2003).

Ray and colleagues accounted for these results by building a *generative model* of agents' behaviour in the Trust Game. Recall that generative models describe processes that are assumed to give rise to some data. With a generative model in hand, one can compute the probability distribution of some quantity that depends on the data. The data, in Ray and colleagues' study, consist in the other agent's decisions about how much money to send. The quantity dependent on such data is the agent's own *decision policy*.

Ray et al.'s model is informed by two facts about agents' cognitive profile. First, people don't have *knowledge* of the outcomes of the alternatives open to them: people may have some expectations about how a certain game may evolve, but they are typically uncertain as to whether such expectations will turn out to be true. Second, people lack knowledge about the *types* of people they are dealing with: people don't know whether others with whom they have only some acquaintance are trustworthy. Ray and colleagues' model assumes that agents have initial beliefs about the type of other agents; all players have prior beliefs about other players' initial type and update their beliefs by implementing Bayesian inference as choices take place.

The types of agents in this model are defined by (i) to what extent they are averse to unequal outcomes and (ii) their level of strategic thinking. The idea of inequality aversion is that people often dislike inequality even when they benefit from the unequal distribution (Fehr and Schmidt 1999). So a type of agent in Ray and colleagues' model is partly defined by how much the agent dislikes disadvantageous inequality (how much she feels envy when somebody else gets a payoff greater than hers) and by how much she dislikes advantageous inequality (how much she feels guilty when she gets a payoff greater than others). The agent's level of strategic thinking together with inequality aversion fully defines the agent's type. Strategizing about what others will do involves thinking about what they think you will do. This sort of thinking can be iterated—so one can think about what others think she thinks others think … and so on. Zero-level players choose completely at random. One-level players think that other players are zero-level, and thereby choose randomly in response to them. Players with two-level strategic thinking think that others are one- and zero- level players, and thereby choose accordingly, and so forth. As explained in chapter 1, people seem to do only a few steps of iterated thinking; usually they just make one step: they decide as though others are choosing randomly. A few make two steps and decide as though others think that they are choosing randomly (Camerer et al. 2004).

On Ray and colleagues' model each agent makes some initial guess about the other agent. The model generates an estimate of what the decisions in the game should be given incoming data: it generates a *decision policy*. Each player can then compare actual and expected decisions and, if the fit is good, infer that her assumptions were probably right. Each player is seeking to maximize her expected pay-off, given their preferences and beliefs about other players' types and her level of strategic thinking.

One of the important features of Ray and colleagues' model is the separation between a "utility (or value) signal" and the signal underlying the inferences generated by the model. As Ray and colleagues (2009) explain, "these distinct signals as to the inner workings of the algorithm […] can be extremely useful to capture neural findings." This separation into distinct signals naturally lends itself to an interpretation in terms of preferences about payoffs in the game and beliefs-dynamics about other agents. The integration of such signals enables the agents to track changes in their social world and behave adaptively. So, given these details, how are beliefs and preferences understood in Ray and colleagues' computational model?

If neurocomputational models such as the one built by Ray and colleagues describe the mechanism (or some aspect of the mechanism) of social decision-making, beliefs and preferences just *are* probability distributions. Social interactions, as typically understood in computational neuroscience, consist in the transmission of messages between agents about their hidden states. Such messages influence the beliefs and preferences of other agents; they affect, that is, the probability distributions encoded by an agent's nervous system. Agents' beliefs and preferences change as they gather more data given rise by the unfolding of the social interaction.

In Ray and colleagues' model, agents' beliefs are probability distributions over the possible types of other players. Agents' preferences are probability distributions over actions given the state obtaining in the world, which in this case is determined by the agent's type and the sequence of plays in the game. If one's beliefs are probability distributions over the possible types of the other agent, then they become more peaked as more observations are made about the other agent's behaviour. One's confidence in some particular hypothesis correspondingly increases. If one' preferences depend on her type, her beliefs about other's types and on the state of the game, and preferences are probability distributions over possible actions, then such distributions become more peaked as more observations are made about the other player's decisions and about the history of the game. One will correspondingly be more likely to select a certain action.

If neural systems deal with uncertainty and encode probability distributions, then my neural representations of your type and of the kind of game we are playing causally affect my neural representation about the utility (or value) of different payoffs distributions in the game. Such neural representations will be used to generate behaviour.

2.1 What Could Neural Representations Be?

Neurons carry information by generating patterns of action potentials, or spikes. Spike patterns carry information about internal and external variables. Cognitive capacities, including the capacity to comply with norms, are enabled by transformations of such patterns of neural activity.

If information is understood in terms of the statistical dependency between a source and receiver (Shannon 1948), then to say that neural spike trains carry information is to say that neural signals are statistically dependent on internal and external variables. Neural signals not only are statistically dependent on some source, but they also *reliably* correlate with their sources: neural signals and the variables with which they correlate seem to constitute a *code*.

"A neural code is a system of rules and mechanisms by which a signal carries information" (deCharms and Zador 2000, p. 614). This code specifies functional relationships between properties of neural activity and properties of internal or external variables. Although it is controversial what the precise rules and mechanisms underlying neural coding are (deCharms and Zador 2000; Dayan and Abbott 2001, Ch. 1), *neural representations* could be individuated as the constituents of the neural code. More precisely, neural representations could be individuated by encoding and decoding mappings between two "alphabets" (Eliasmith 2003). And neural representing could be described as a two-stage encoding and decoding process.

Neural *encoding* refers to the mapping of some variable *s* onto the response of one or more neurons, **r**. It specifies the functional dependence of some neural property on some property of a stimulus. Action potentials are the basic units of the encoding alphabet. Neural *decoding* refers to the estimation of some property of some stimulus from some property of some neural response. It specifies how some value of some physical variable *s* can be readout from a neural response pattern **r**. Physical properties are plausibly the basic units of the decoding alphabet. The estimate $\hat{s}$ yielded by the decoder is used by the system to generate behaviour.

To get to grips with the concept of neural representation as encoding-decoding mappings, consider perceptual visual beliefs. Visual neurons code physical properties with their activity in response to stimuli. The action-potential firing rates of neurons in the primary visual cortex reliably co-varies with and selectively responds to properties such as spatial location, orientation, and direction of motion of visual stimuli (Hubel and Wiesel 1962). Neural encoding provides a mapping from stimulus to neural response. Given a stimulus, neural encoding determines how neural activation in a certain brain area transduces the stimulus in function of some non-neural variable or parameter. The standard tool to describe how neural activity depends on some physical property is the neural *tuning curve*: a plot of the average firing rate of the neuron in function of relevant stimulus values.

Neural decoding provides a mapping from neural response to stimulus. From neural tuning curves, it is possible to extract an estimate of which property is coded by a particular neural activation. The tuning curve to a feature of a stimulus—e.g. the orientation of a bar—is the curve describing the average response of a neuron in function of the values of the feature. A decoder determines how the information carried by neural activity population is used by the rest of the system. Given a certain neural activation, the study of neural decoding amounts to estimating how likely it is that a certain stimulus is in the environment—e.g. amounts to determining the orientation of a bar of light given a pattern of activation in the primary visual cortex. The decoding corresponds to the task performed by neurons downstream when they read off the spike trains that are their inputs.

As Chapter 1 suggested, the neural code might comprise a "representational hierarchy": complex, abstract representations might be encoded at higher levels in the hierarchy computed in function of low-level representations *and* some generative model. Transformations of some variable *s* to certain behavioural responses or internal changes—possibly driven by Bayesian inference—might be implemented along an encoding-decoding cascade. Provided that structural and mathematical relationships between levels in the hierarchy are defined, it might be possible to systematically relate higher-level neural representations to their lower level components (Cf. Eliasmith 2003).

In light of Ray et al's (2009) work, when agents play a trust game, some of their high-level neural representations carry information about other players' types, some carry information about the action to implement given the current state of the game. With their concerted activations and transformations, these neural representations lead to adaptive behaviour in response to other agents' behaviour. Provided one player's pattern of neural response, neural decoding can yield probabilities for each value of *s*—where *s* spans over types of opponents—that such value has led to the observed firing pattern **r**, and then selects one appropriate value $\hat{s}$. How exactly the decoder estimates the stimulus that has led to a certain firing pattern and which value $\hat{s}$ it yields depend on the information available to the system (e.g. on the detail of the generative model it could use) and on how the overall estimation errors are weighted (i.e. the type of loss function used by the system).

For example, under conditions of minimal information, if *s* is a parameter spanning over types of opponents, **r** is a spike train, and we *only* know the encoding mapping *Prob* (**r**|*s*), then *one* possible way to readout the spike trains is by means of maximum-likelihood decoding. The maximum-likelihood estimate $\hat{s}$ is the stimulus that has maximal probability of having caused the response **r**, that is $\hat{s}$ = argmax_s *Prob* (**r**|s).

Note that a probabilistic way to characterize the decoder underwrites the fact that neurons are noisy, have graded responses to stimuli, and that might also encode the uncertainty associated with a stimulus with their firing rates. In a sense, it is misleading then to say that neurons are detectors that determine, for example, that either one agent is trustworthy or not: "Neurons don't 'detect' things (i.e. they don't determine that there *is* an edge or there *isn't* one), they respond selectively to input, the more similar the input, the more similar the response" (Eliasmith 2005, p. 118).

Having hinted at what neural representations could be, I now argue that norm compliance should be explained by appealing to representations; then I turn to argue that norm compliance should be explained by appealing to *neural* representations.

3. The Indispensability of (Neural) Representation. Or "Representational Hunger" Strikes Again!

The argument for why norm compliance should be explained by appeal to representations has two premises.

P1. Internal representations give us unique explanatory leverage regarding agents' behaviour in "representational-hungry" problem domains.

P2. Paradigm cases of social norm compliance consist in behaviour in "representational-hungry" problem domains.

C. Therefore internal representations give us unique explanatory leverage regarding paradigm cases of social norm compliance.

The argument is deductively valid. Premise 1 involves the notion of "representational hungry" problem domain. This notion is elaborated by Clark and Toribio (1994). As Clark and Toribio define it, a problem domain is "representational-hungry" just in case "one or both of the following conditions apply:

1. The problem involves reasoning about absent, non-existent, or counterfactual states of affairs.

2. The problem requires the agent to be selectively sensitive to parameters whose ambient physical manifestations are complex and unruly (for example, open-endedly disjunctive)" (Ibid., p. 419).

Clark and Toribio argue persuasively—I think—that internal stands-in, or representations, are necessary to successfully tackle representational-hungry problem domains, and hence representations give us unique explanatory leverage regarding agents' behaviour in such domains. Here I take P1 for granted, and focus on P2 which is more controversial. I argue that conditions 1 and 2 apply to paradigm cases of norm compliance, and therefore paradigm cases of norm compliance consist in behaviour in "representational-hungry" problem domains. As paradigm cases I focus on the type of trust game modeled by Ray and colleagues, and on a more ordinary case. If norm compliance paradigmatically takes place in representation-hungry problem domains, then internal representations give us unique explanatory leverage regarding paradigm cases of social norm compliance. Let's consider the trust game.

To trust someone implies some degree of uncertainty: You take the risk of betrayal. You repay another's trust even though that may go against your interest to maximize your profit. When you trust strangers you don't know whether they are motivated only by a selfish desire to maximize their own profit. Finding out the type and beliefs of other agents in a trust game requires an ability to *anticipate* their actions and to reason *counterfactually*. Condition 1 then applies to this case. Anticipation and counterfactual reasoning, as argued by Clark and Toribio, seem to require the use of inner resources. Ray and colleagues' generative model is one type of inner resource which enables an agent to behave appropriately even in absence of explicit inputs specific to other players' type. Hence Trust Game- types of situations are "representational-hungry."

Consider this other situation. There is this social norm in football: When a player goes down injured, the ball is usually kicked out of play to allow the player to receive treatment. If the ball is kicked out of play by the opponents, a further norm is to return the ball to them. These norms have never been formalized in the rules of the game, but furious reactions are likely to ensue if somebody fails to comply with them.

Imagine now that you are playing an important football game. You notice that a football player from the opponent team looks as though he is injured. You have the ball and you can set up a team-mate for a goal. Should you put the ball out of play? The decision to pass the ball to your team mate or to throw it out to allow the opponent player to receive treatment takes fractions of seconds. It is very likely to be unconscious and driven by heuristics. Nonetheless, *counterfactual* reasoning and *anticipation* seem to play an important role in this occasion as well. You need make a rapid judgement concerning the actual state of the opponent. You need find out whether the opponent is in fact injured. You need judge what could happen if you played on and the opponent was in fact injured; you need anticipate the reactions of the opponents if he failed to put the ball out of play. Therefore, abilities for counterfactual reasoning and anticipation—if probably unconscious and driven by heuristics—seem essential to your decision to comply with the norm.

The same problem domain in football requires that the player who is to make a decision is "selectively sensitive to parameters whose ambient physical manifestations are complex and unruly" (Ibid.). Imagine that the ball has been kicked out of play because a player went down injured. It is time for a throw-in. It is known that you give the ball back, if an opponent player deliberately kicked the ball out of play because a team-mate of yours was injured. One condition for the player to comply with this norm is that he is sensitive to abstract, relational properties such as the value of "fair-play," "reciprocity," or "cheating." The physical manifestation of such relational properties as "fair-play" is typically "complex" and "unruly," since in general whether a pattern of physical features in a social situation counts as "fair-play" depends on other features obtaining or failing to obtain in that situation, and on the learning trajectory of the agents involved. So, people do not seem to rely on invariant general rules when they need to identify a certain pattern as "fair-play."

In order to track such types of properties, one needs to rely on internal representations. Given his previous experience in the world of football, the football player has developed a capacity to track those abstract properties across situations. Clark (2000b) calls this capacity *representational re-coding*, whereby complex, abstract relations are re-coded into simple, usable objects—more on representational recoding in Chapter 5. Given a diverse array of perceptual inputs, courtesy of representational re-coding one can compress that array into an item whose content corresponds to an abstract property. The item can be stored in memory and retrieved for further processing without the need to store and retrieve all of the diverse perceptual inputs underlying it. Before the throw-in under those circumstances the player's sensitivity to such properties as "fair-play" is important to explain his behaviour. Such sensitivity depends on representational re-coding. Since the idea of internal representation is essential to this kind of re-coding, it follows that the idea of internal representation is essential to explaining the player's behaviour. The problem domain that our football player faces is an instance of "representational-hungry" problem.

If my accounts of the Trust Game and of two social norms of fair-play in football are correct, then paradigm cases of social norm compliance consist in behaviour in "representational-hungry" problem domains (P2). It follows from the argument stated at the beginning of this section that internal representations give us unique explanatory leverage regarding paradigm cases of social norm compliance.

4. Representational-Hungry? On Dreyfus's Anti-Representationalist

Contrary to what I just argued, according to Hubert Dreyfus, paradigm cases of social norm compliance do *not* consist in behaviour in "representational-hungry" problem domains.

Dreyfus claims that "*some* central cases of intelligent behavior do not involve mental representation" (Dreyfus 2002b, p. 414). And social norm compliance, most of the time, falls among those "central cases of intelligent behaviour." Paradigmatic cases of social norm compliance, for Dreyfus, consist in *non*-representational hungry behaviour. If Dreyfus is right, then the argument laid down in section 3.1 is unsound.

Dreyfus (Ibid., p. 417-418) asks us to consider a situation in the elevator. The elevator stops at the seventh floor and two people step in. The people already in the elevator shuffle and move around until they are at appropriate distance from the others. This is a paradigmatic case of social norm compliance. According to Dreyfus, the situation just described is not hungry for representation. Rather, it is an instance of "skillful coping" which amounts to a spontaneous responsiveness to the demands of a situation. Skillful coping does not require either deliberation or attention, and importantly does not involve the representation of goals. If norm compliance is typically an instance of "absorbed skillful coping," then we need and should *not* explain norm compliance with recourse to representations.

Dreyfus draws on Merleau-Ponty's work to account for paradigm cases of norm compliance in terms of the *intentional arc* and the tendency to achieve *maximal* *grip*. In order to give flesh to these two notions, he borrows from certain features of neural networks modelling and from Walter Freeman's (1991) attractor theory of the brain dynamics underlying perception and action. For Dreyfus, "neural networks exhibit crucial structural features of the intentional arc," and Freeman's account might underlie maximal grip (Dreyfus 2002a, p. 413). The resulting explanatory framework is one where representation plays no role.

Dreyfus's argument assumes a particular concept of representation, which is not the one I put forward. Now, after having introduced the notions of the intentional arc and of maximal grip, I explain why representational hungry domains need not involve representations as conceived of by Dreyfus.

4.1 Representations After All?

"The *intentional arc*—Dreyfus explains—names the tight connection between body and world" (2002a, p. 367). The intentional arc describes a relationship between agent's skills and the world: when agents acquire a skill, becoming experts in doing something, the skill manifests itself spontaneously given certain solicitations of a situation. The intentional arc does not depend on representations stored in the head: skills underlain by the intentional arc are finer and finer dispositions to respond to cues in the world. This kind of body-world relationship grows via extensive interaction with other agents and by "dealing with things and situations."

"*Maximal grip* names the body's tendency to respond to these solicitations in such a way as to bring the current situation closer to the agent's sense of an optimal gestalt" (Ibid., pp. 367-368). Maximal grip describes the process whereby the agent comes to "see" how to be drawn by environmental solicitations to realize a particular goal without representing the goal. Dreyfus argues that the way neural networks learn vindicates the notion of intentional arc, and that Freeman's attractor theory might be one way to flesh out the sub-personal mechanism of getting a maximal grip on a situation.

Dreyfus's reliance on neural networks and dynamical system theory is indicative of how he conceives of representation. Dreyfus associates the notion of representation with the "classicist" idea of strings of symbols tokened in a system, which are isomorphic to *propositional attitudes* (e.g. Fodor and Pylyshyn 1988; Newell and Simon 1972). But such data structures are only one way of understanding representation. Representational hungry problem domains need not require *this* type of data structure.

Both connectionists and Freeman (1991) speak in fact of representations although they don't have in mind the classicist notion. Andy Clark (2002b) commenting on Dreyfus raises exactly this worry: attractor states in dynamical systems and high-dimensional weight spaces of neural networks can be understood "as new powerful kinds of internal representations" (p. 386). The way I characterized neural representations in terms of encoding-decoding mappings fits with the way neural networks learn and with the way brain uses attractors. If this is so, then social situations like the one in the elevator described by Dreyfus can still be hungry for representations, although non-classicist representations. In order to establish this point, I consider Dreyfus's response to Clark's worry.

Dreyfus (2002b) has two complaints. He claims that the use of representation in neural networks and dynamical system theory is unwarranted. In those contexts the notion of representation is too weak "to do the job of showing that *particular* brain states are correlated with *particular* items in the world, let alone that they have content, that is, that they *represent* such particular items under an aspect" (p. 420). The first complaint has to do with the quality of the correlation between neural activation and physical features in the world. The second has to do with how neural activations can represent external stimuli under an aspect—e.g. seeing a carrot under the aspect nourishment.

A characterization of neural representation in terms of encoding-decoding provides us with an answer to Dreyfus's first concern. The input to a neural network is encoded by a certain pattern of activation. From a given neural activation, the system decodes information about the input. Although neural networks do not store particular rules for dealing with particular inputs, they give the same or similar outputs to same or similar inputs after training. Encoding-decoding mappings can be formalized as probability distributions, which can reliably specify correlations between *particular* neural activations and *particular* physical features in the world. Hence, the notion of representation I put forward is strong enough "to show that particular brain states are correlated with particular items in the world" (Ibid.). The same argument applies in the context of brain dynamics.

Dreyfus (2002b, p. 420) recognizes that Freeman himself claims that the brain uses attractors to *represent* causes in the sensorium (but see Freeman and Skarda 1990). Dreyfus, however, asks us to resist representation-talk in this case. He points out that "when the rabbit smells and successfully eats a carrot, it forms a new attractor, and that attractor, in an appropriate context, will henceforth cause the rabbit to go for a carrot, this is just a complex physical event" (Ibid.). But here Dreyfus is describing an example where a particular brain state may be correlated with a particular feature of the world. At a minimum, the attractor in the rabbit's brain is capable to stand-in for the carrot, say when the carrot is not here and now, the rabbit is hungry and is directed towards carrots. In light of this, "[w]hat makes one want to use representation talk" is *not* as Dreyfus's claim "that the complex event of the system relaxing into an attractor basin is isomorphic with the agent's experience of being drawn towards an equilibrium" (Ibid.). Rather, it is the fact that appeal to representation is justified, minimally, when an entity stands-in for some possible state of affairs. Since the attractor in the rabbit's brain stands-in for the carrot *and* is consumed by the system via decoding processes, we want to use representation talk also in the context of systems dynamics.

Dreyfus's second complaint has to do with content. He wonders how representations in neural networks and system dynamics can represent "particular items under an aspect." For example, can the attractor in the rabbit's brain represent carrots under the aspect nourishment? Understanding neural representations as encoding-decoding mappings suggests one way in which the attractor in the rabbit's brain does represent carrots *as* nourishment. Decoding determines the relevance of the encoding for the system. It specifies how neural activations are used by the system to produce behaviour. Particular activations decode certain features of the world in a larger system of encodings and decodings. Within this larger system, along an encoding-decoding cascade the representation of the carrot is probably associated with the representation of a high-level property like nourishment. Properties such as edible and dangerous may also be encoded in neural activity. Encodings of such properties might depend on encodings-decodings of "low-level" properties like "displacement," "mass," and "orientation" (Eliasmith 2003, p. 502).

There is no reason why the attractor in the rabbit's brain cannot represent carrots under the aspect nourishment. Representation as encoding-decoding, therefore, might also do the job of showing that particular items are represented under an aspect.

Dreyfus has not established that the mechanisms that might underlie the intentional arc and maximal grip are representation-free. *Even if* the intentional arc and maximal grip are in place in situations where people comply with norms, those situations can still be representational-hungry. Let me expand on this point by reconsidering the situation in the elevator where a person steps in.

4.2 Shuffling in the Elevator. Systemic Dynamics and Causal Couplings

Why do people in the elevator shuffle until they get to an appropriate distance? For Dreyfus this is an example of "spontaneous absorbed coping." Dreyfus's explanation is that after repeated interaction with others in elevators, people have acquired a disposition to respond to the solicitations of that kind of situations by getting to an appropriate distance. Nobody can specify that distance. Nobody is trying to get to that distance. People in the elevator are drawn to get there by responding to the whole [elevator-person A-person B- person C- etc] situation. They do not respond to particular features. They do not represent the person who is stepping in as a separate feature of the situation. "The embodied agent—Dreyfus explains (2002b, p. 420)—doesn't *think of* doing what is solicited either. He just let himself be drawn to lower a tension and straightway finds his body doing what feels appropriate, without needing to, or being able to, represent some desired goal." Dreyfus's explanation is couched in terms of a perception-based fine-grained disposition in an extended body-environment system. This explanatory framework has two features: (i) embodied agents respond to the *whole* situation, (ii) embodied agents coping with their situation do not have any representation of their *goal*.

The main motivation for (i) is causal coupling. That agents are coupled with their surroundings means that the agents continuously both affect and are affected by what surrounds them. Coupling is usually taken as reason in support of the arbitrariness of distinguishing brain-centered cognitive systems from the environment where they are embedded (Beer 2008). Causal coupling would constitute a reason to doubt that the situation in the elevator is representational-hungry.

To understand the interactive complexity underlying skilful coping—runs Dreyfus's argument—we should adopt a "wholist" perspective. According to this argument, the situation in the elevator is best explained in terms of dynamics of the whole system [elevator-person A stepping in-other persons in the elevator] evolving towards an adaptive equilibrium. The bottom line is that paradigmatic cases of norm compliance may not involve either a behavioural or a neural ability, but systemic dynamics. If they essentially involve systemic dynamics, then it is mistaken to view such situations as involving specific representational components.

However, in cases of skillful coping we still have good, *independent* reason to ask about information-processing components representing specific features of a situation. A brain mechanism underlying the capacity to respond to certain external solicitations—e.g. a person stepping in the elevator—is taken to be coupled to the whole body-environment because we have a representational pre-understanding of its role: we have a pre-understanding of the type of information the mechanism could carry and manipulate. Without this kind of understanding it would be problematic to identify where to apply the dynamicist analysis—whether at the level of brain-body-environment system, or of body-neuromechanical interactions, or neural interactions. For there wouldn't be an *independent* rationale to understanding why we should (de)couple possible components of the system in certain ways rather than others.

The second feature of Dreyfus's account of the elevator case is (ii) that the embodied agent coping with her situation does not have any representation of her *goal*. Dreyfus starts with a puzzle. During skill-acquisition agents modify their behaviour in function of their results. When the action results in a failure, then something needs to be revised. But in order to adjust one's behaviour in function of failure and success, some representation of a goal seems to be necessary. Such a representation specifies a target-state that determines appropriate adjustment in the agent's behaviour. If this is so, then it seems that all skilful action requires goal-representation. If one is acting skillfully, then there is something she is trying to do. If there is something she is trying to do, then she is pursuing a goal. Therefore, goal-representations seem to be necessary for skillful action.

Dreyfus resolves the puzzle by rejecting the first conditional. It is not always the case that if one is acting skillfully, there is something she is trying to do. As mentioned above, Dreyfus claims that "[i]n general, we don't have to *try* to comport ourselves in socially acceptable ways" (Dreyfus 2002b, p. 418). His argument is that we experience such kinds of situations "as drawing the movements out of us" (Dreyfus 2002a, p. 380). In "the *experience of acting*" the direction of causation is not from a represented goal to the world. It is the world itself that initiates certain of our bodily movements drawing us towards appropriate actions: No goal-state is pursued in norm compliance. Success in complying with a social norm is assessed as experience of *optimal gestalt*.

In essence, Dreyfus's argument is this: For some skillful action like some cases of norm compliance we experience the situation as drawing the appropriate action out of us. If this is so, then for some skillful action we do not experience our goals as causing our action. We experience that the direction of causation goes from the situation to the action itself. Hence we do not experience our goals. Therefore, the representation of goals is not involved in some skillful actions like certain instances of norm compliance.

I think this argument is a *non-sequitur*. Assume that we do not *experience* norm compliance as caused by the pursuit of a goal. Assume also that sometimes we cannot *formulate* the goal that we may pursue in certain contexts. For example, we cannot tell what is the socially appropriate distance to maintain in an elevator. From these, it does not follow that the representation of a goal is not involved in norm-compliance. It only follows that the representation of a goal in certain instances of norm compliance is not explicit and conscious. In such cases, the representation of the goal may be tacit, unconscious, or dispositional as characterized in Section 1.

The explanatory leverage given by goal-representations in the case of norm compliance has to do with both the anticipatory and evaluative nature of goals. On the one hand, goals indicate potential future states of affairs towards which we are driven. They govern our behaviour towards the realization of that state. On the other hand, goals indicate valuable states of affairs. They allow us to evaluate the current state of affairs in function of the target-state. When a person steps in a crowded elevator, goal-representations provide us with a natural explanation of why people start to shuffle until they reach a certain position. Each agent may have the goal of keeping a socially appropriate distance from the others. The current state is confronted with that goal. If the state fails to fit the goal, a prediction-error ensues and some adjustment is required. The goal-representation enables the agent both to anticipate what might happen if the target state fails to be reached and to evaluate that certain possible states are "bad" whereas others are "good."

5. Explanatory Virtues of Neural Representations

Ramsey (2007) argues that it is always *possible* to treat a system as representational, but it is never necessary. Ramsey puts forward the challenge to specify what the positing of representations could give us in terms of non-trivial explanatory purchase. I now address this challenge by comparing representationalism with dispositionalism and focusing on the manipulation and control afforded by neural representations over behaviour.

Suppose that dispositionalism is the right way to think about belief and desire. Suppose that beliefs and desires understood as behavioural dispositions enter an explanatory, causal relationship with norm compliance behaviour. What is the type of control and manipulations of norm compliance behaviour that such a relationship can facilitate? To what extent understanding beliefs and desires as behavioural dispositions facilitates us to control, manipulate, and predict norm compliance?

There are two distinct questions here. The first has to do with metaphysics and asks whether beliefs *as* behavioural dispositions can be causes. The second has to do with explanation and asks whether beliefs understood as behavioural dispositions carry extra explanatory value in a complete causal explanation of norm-compliance behaviour. After some brief considerations on the first issue, my focus will be on the second question which is related to Ramsey's challenge.

5.1 Dispositions as Causes?

Let's distinguish between dispositions and their categorical bases. Fragility, irascibility, and perhaps expectations are examples of dispositions. The categorical basis of the fragility of a glass is its physico-chemical structure. If expectations are dispositions, then their categorical bases are certain brain states. Granted such a distinction, the argument for why dispositions cannot cause behaviour is analogous to Kim's causal exclusion argument about mental causation (Kim 1998). If all physical effects have sufficient physical causes and no physical effects are caused twice (that is, there are no overdetermining causes) by distinct categorical and dispositional causes, then there cannot be dispositional causes.

Pursuing this line of argument, Prior, Pargetter and Jackson (1982) argued that dispositions do nothing. It is the categorical basis of a disposition that causes things. If this argument is sound and if beliefs are just dispositions, then beliefs do nothing. In one note Schwitzgebel (2002, note 18, p. 273)—who describes and defends a dispositional account of belief—acknowledges that one may be concerned that a dispositionalist view "doesn't allow for beliefs to cause behaviour." He suggests that one way to deal with the problem is to identify "believing with being in a certain categorical state." Hence belief would cause behaviour. However, by following this strategy, it seems that believing will have more to do with having the right kind of internal categorical basis than with being disposed to do certain kinds of things. Still, even if dispositions are not causes, and hence beliefs understood as dispositions are not causes, it does not follow that beliefs as dispositions would not give us some leverage in causal *explanations*.

5.2 Manipulation and Control. Representations' Explanatory Purchase

There is good evidence that social norm compliance can be affected by what an agent expects others would do in a similar situation. An agent's tendency towards norm compliance is also affected by other kinds of expectations: by what one believes others think she ought to do in that type of situation. Agents' beliefs and expectations in some social situation can be manipulated by providing them with information about other agents' judgements and behaviour in the same type of situation. By manipulating their beliefs and expectations, agents' tendency to comply with a given norm in that situation can change (e.g. Bicchieri and Xiao 2009).

Assume that belief and preference are just dispositions to behave in certain ways under appropriate circumstances. Suppose that in a Trust Game the information provided to the players causally affect their disposition to reciprocate. Since preferences are dependent on beliefs in Ray et al.'s (2009) model[3] (recall that in their model beliefs about your type influences my preferences about payoffs in the game), the provision of a certain type of information about the type of trustee causes the investor to be disposed to prefer, for example, to invest nothing. The investor's beliefs are manipulated by the provision of a certain type of information, which affects the investor's preferences about payoffs. If to prefer something to something

[3] Also in Bicchieri's (2006) account of norm compliance preferences are dependent on beliefs. On her model, an agent's preferences are conditional on his or her own beliefs regarding other people's actions and expectations. So one *prefers* to follow a norm if he or she *believes* that certain conditions occur.

else is just the disposition to do what realizes the former thing rather than the latter, we should say that information about the trustee's type causes the investor "to be disposed to have a disposition" to invest nothing. This last expression sounds strange, as it seems that one cannot be "disposed to have a disposition." But on a dispositionalist understanding of beliefs and preferences that is the way we should explain the investor's decision to invest nothing under certain circumstances. The investor's preferences would be second-order mental states elicited by beliefs. Here is a possible explanation of the investor's behaviour.

The investor prefers to invest nothing in the game *because* she expects that the trustee will not reciprocate. She expects that the trustee will not reciprocate *because* she has received a certain type of information about her type. The 'because' is causal in both statements. The first 'because' connects two dispositions: an expectation and a preference. The second connects a type of disposition *viz.* an expectation, and a piece of information. In the second statement we can individuate the cause as a physical process *viz.* the transmission of messages about the trustee's history of plays. This way of individuating the cause enables us to manipulate it: for example we can destroy the message before it reaches the investor or we can modify it by adding noise.

There are two questions that this explanation leaves unanswered. *Why*, or in virtue of what, does that message about the history of the game cause the investor to have a certain expectation? *Why*, or in virtue of what, does the expectation cause the investor to have a certain preference? The answers to these questions are important if we want to intervene causally on the investor's expectations and preferences.

To make the point stick, think about this question: “Why did your mug break when Courtney dropped it?” You can answer: “Because it was fragile and Courtney dropped it.” The problem with this explanation is analogous to that of the explanation above: it doesn’t tell us what we should do if we wanted to prevent the mug from breaking when dropped. It doesn’t facilitate us to individuate how we should intervene if we wanted to manipulate or control the effects of dropping the mug. Another possible answer to the question above is: “The mug broke because it has such and such atomic structure and Courtney dropped it.” This explanation places us in a better position to manipulate and control the effects of dropping the mug. For example, we can manipulate the atomic structure of the mug in order to control the effects of dropping the mug. There is no appeal to fragility here. Still we have provided a satisfactory explanation that can *also* facilitate manipulations and predictions regarding the behaviour of the mug.

It may be complained that routinely when ordinary people wish to control the effects of dropping a mug they intervene on fragility by protecting the mug with some packaging material. People don’t intervene on its micro-structure. Hence, when it comes to intervention the disposition-free explanation given above is irrelevant. This complaint, however, is misguided.

Mugs and other fragile objects are ordinarily protected with packaging material when they are shipped or transported. Under those circumstances we cannot use mugs as drinking cups. We have to unwrap them to use them as drinking cups. Mugs routinely break when they are used as drinking cups. Thus, if we intervene on the fragility of the cup by wrapping it with packaging material, the mug is useless as a drinking cup. That kind of intervention would prevent us from using the mug as a mug. Instead, by intervening directly on the micro-structure of mugs, not only they could be shipped safely, they could also be used as drinking cups. The bottom line is that by explaining something we want to understand mechanisms. One of the reasons we want to understand mechanisms is that we should intervene on mechanisms, if we wish to control and manipulate effectively certain phenomena.

We ascribe dispositions as global properties of a whole system. So, the fragility of a glass is not localized in any distinct part of the glass. A glass is fragile throughout all its parts. The atomic microstructure of the glass instead can be inhomogeneous. It is because of the microstructure of the glass that we can say that the stem of a wine glass is more fragile than the bowl. By individuating *where* the glass has a certain structure, we can say that that part is more fragile than another *and* we can intervene locally on that part. To say that beliefs and desires are global dispositions of a whole system our belief-desire explanations would not facilitate us to identify *where* we should intervene in a cognitive system to make a difference in its behaviour. This is not to claim that beliefs and desires must be localized in some particular part of the system. To have a belief or a desire is to have some neural representation which typically arises from the activity of distributed, and yet *identifiable*, populations of neurons.

If we wanted to intervene causally on an agent's mental states, would an explanation couched in terms of neural representations be a better guide than an explanation couched in terms of dispositions? Ray and colleagues' case helps us answer this question in an affirmative way. The social utility function that they implemented represents the agent's mental states in a way that "mandate probing, belief manipulation and the like" (Ray et al. 2009). An explanation couched in terms of dispositions has difficulties, in comparison to an explanation couched in terms of neural representations, to provide us with an answer to the question *why*, or in virtue of what, a player manipulates another player's beliefs. The 'why' here is causal. An answer to this question will enable us to individuate where and how to intervene to cause certain effects.

Suppose that two human subjects are playing a Trust Game. Some neural representations encode preferences in the form of utility (or value) signals. Other neural representations encode beliefs; neural computations might underlie belief-dynamics in the form of inferential schemes embedded in a generative model. Assume that we appeal to these neural representations to explain why a player is playing fair. The two separate signals of utility and inference in the algorithm underlying Ray and colleagues' model map onto the activity of identifiable neural populations. Suppose that Ray and colleagues' model enables us to estimate in real time brain activity and to decode the information therein represented. The information encoded in neural representations can then be extracted and manipulated by altering some parameters of the algorithm carried out by the system. After manipulation the information can be fed back to the players' brain courtesy of appropriate techniques. If successful, the manipulation would lead to changes in brain activity and behaviour.

Research in computational neuroscience and brain machine interface is beginning to make the scenario just described less science fiction than it can seem. Kawato (2008a) illustrates how the combination of computational models, brain-network interfaces –which non-invasively estimate neural activity and read out the information carried by neural activity—and decoding algorithms can foster what he calls *manipulative neuroscience*.

Kawato (2008b) reports on a project where a monkey's brain activity could control a humanoid robot across the Pacific Ocean. In this project the pattern of activity of certain populations of neurons encoded in a monkey's motor cortex were recorded while the monkey was engaging in a motor task in a lab in the United States. The kinematic features of the monkey's motions were decoded from neural firing rates and sent via an internet connection in real time to a robot located in Japan. Courtesy of this signal, the robot could execute locomotion-like movements similar to those performed by the monkey. Another instance of manipulative neuroscience is the remote radio control of insect flight. Sato and Maharbiz (2010) review studies where insects in free flight are controlled courtesy of implantable interfaces. Courtesy of an implant for neural stimulation of an insect's brain coupled with low power radio systems, the insect can be put into motion, stopped and controlled while it is in flight. In light of this type of research, manipulative neuroscience "has already moved beyond mere science-fiction fantasy in the domain of sensory reconstruction and central control repair as exemplified by artificial cochlear and deep brain stimulation" (Kawato 2008b, p. 139). It does not seem to be a mere whim of fantasy to expect that non-trivial choice behaviour in social contexts might be manipulated in similar ways. One of the reasons behind this type of research is to show that we do understand some aspects of behaviour well enough to carefully manipulate it.

The notion of a neural representation, it seems, yields non-trivial understanding here. To begin with, if behaviour depends on generative models, then agents rely on *assumptions*—that is, on representations—about how cues concerning e.g. other agents' types and decisions are generated. Agents transform these assumptions so as to determine which behaviour they should implement if they want to behave adaptively. For example, if the goal of the nervous system of agents playing the Trust game is to estimate the hidden variable "opponent's type," then, assuming that a particular opponent's type generated the observed cues, the agent has to invert her generative model, and estimate the hidden variable other player's type by combining the cues she observed.

Secondly, all the successful cases of manipulative neuroscience involve some aspect of the notion of a neural representation. Manipulations and control of agents' behaviour, in fact, leverages encoding-decoding mappings between a neural alphabet and a physical alphabet. Identification of neural representations enables one to dissect them into components at lower-levels or to recombine them in ways sensitive to the information they carry. Furthermore, identification of neural representations could facilitate us to guide agents' behaviour in the absence of the properties those representations are about. The notion of a neural representation, therefore, seems to be necessary to all successful cases of manipulative neuroscience.

Explanations of norm compliance couched in terms of neural representations, in comparison to explanations couched in terms of dispositions, may facilitate the *direct* manipulation of information carried by neural activity. They may provide information about *where* one should intervene in order to cause certain effects. Although the appeal to neural representation may lead one to think that the only types of manipulation facilitated by this framework are neurobiological, this is not the case. Emphasis is put on the informational content of neural activation. If we had a better understanding of the precise nature of the neural code, there would be more reliable grounds to identify what kind of information causes certain changes in neural activations, which ultimately causes one to comply with a norm.

Conclusion

In so far as computationalism is bound up with representation, this chapter has provided independent reason in support of a neurocomputational approach to explanation of norm compliance. After having distinguished between dispositionalism and representationalism as ways to understand what it is to have a belief or a preference, the chapter has argued that the explanation of paradigmatic cases of norm compliance requires the appeal to neural representations. Furthermore, if we wish to control and manipulate effectively behaviours like norm compliance, beliefs and desires are better understood as neural representations instead of behavioural dispositions.

CHAPTER 3.

On the Representational Format of Social Norms: A Map

If the mechanism of norm compliance implements inductive inferences and deals with representations, then a question to be addressed for a Bayesian-RL model of norm compliance is this: What is the format of the background knowledge[4] that supports such inferences? This chapter explores this issue. Specifically, it addresses how social representations, on which norm compliance would depend, could be stored in memory. This topic is important because a thorough assessment of arguments for or against a given account of the mechanism of social norm compliance requires that we have some grip on the representational format in which social representations could be stored in memory.

The aims of this chapter are neither to develop a theory of concepts nor to explain how exactly the content and identity of social representation are fixed. Some of the studies I consider appeal to the notion of "concept" (Murphy 2002 provides a review on the psychology of concepts), but I do not appeal to such a notion. I limit myself to the notion of a social (neural) representation as characterized in Chapters 1 and 2. This is because the appeal to the notion of "concept" may engender confusion and give rise to problematic issues, like the issue of content invariance, which go beyond my aims here.

The aims of this chapter are twofold. First, the chapter aims at drawing a map of the main options for the format of social representations. Second, starting from the assumption that social representations need not be in one single format, the chapter urges that a probabilistic approach to social cognition is especially fruitful for

[4] My use of 'knowledge' here is akin to the cognitive scientists' use, as "body of information." This use is noncommittal to truth or justification.

evaluating proposals about the different forms that social knowledge can take across different domains and tasks.

The chapter is in three sections. The first section begins by focusing on one of the functions of social representations, namely: categorization. The cognitive science of categorization will help me to sketch a map of the main possibilities to account for how social representations may be stored.

Note that I do *not* consider the view that social representations are stored as rules in a sentence-like format possibly regimented with deontic logic. This is for two reasons. First, the relationships between social norm compliance and linguaform rules will be considered in Chapters 4 and 5. Second, philosophers have lavished a great deal of attention on the relationship between rules and moral thought, while they have overlooked other possible formats in which moral knowledge can be represented (Stich 1993). Here, I consider three such alternative formats: prototypes, exemplars and scripts (for a similar, more nuanced map of "concepts" see Machery 2009, Ch. 5).

The second section draws some of the empirical consequences of each alternative, and motivates what type of evidence could count for or against them. Although I don't assess the evidence in greater detail, the last section notes that much of the contemporary research in the cognitive science of categorization assumes that there is telling evidence that we use representations stored in multiple formats.

Starting from this assumption—namely that social knowledge relevant to social norm compliance may well be stored in multiple representational formats—the last section presents a general probabilistic approach that can be useful to explore how inductive inferences underlying social categorization and norm compliance can draw on knowledge in multiple formats.

1. How Can Social Representations Be Stored? Three Formats

The Godfather (1972, directed by Francis Ford Coppola) opens with a puzzling situation. Don Vito Corleone (Marlon Brando) is listening to pleas for favors in his office, while guests are celebrating his daughter's wedding reception in the sunny outdoor veranda. Don Vito's behaviour can appear to be socially inappropriate: Why is he not partying with her daughter and the other guests? Tom Hagen (Robert Duvall), family lawyer and Don Vito's "consigliere," explains: "It's part of the wedding. No Sicilian can refuse any request on his daughter's wedding day." Tom appeals to a social norm in order to explain Don Vito's behaviour.

Don Vito's case illustrates one interesting aspect of social norms. In order to make sense of a social situation so that we can see what type of behaviour is appropriate or inappropriate in it, we rely on categorization. The activation of social norms requires that we categorize events, individuals, and objects in some specific ways. To make sense of Don Vito's behaviour, a social situation has to be understood *as* a wedding reception, Don Vito must be seen *as* a Sicilian "padrino" (as a Sicilian godfather), and so forth.

Categorization can employ social representations. Social representations function as categorization devices by enabling the agent who possesses them to assign instances of events, individuals, objects and situations to their categories, and to make inferences about newly encountered events, individuals, objects and situations on the basis of stored knowledge about those categories. In what follows the expressions ‘social representations’ and ‘categorical knowledge underlying norm compliance’ are used interchangeably.

The problem of categorization is to classify an item (e.g. some event, individual or object) as belonging to a particular category. Categories are knowledge structures corresponding to non-arbitrary classes of events, individuals or objects. This classification can be used in different ways for different purposes. Drawing upon our categorization of an item as belonging to a certain category, we can infer, for example, unobserved properties of the item based on common properties within the category. Socially appropriate behaviour requires an ability to learn social categories and to use them correctly across situations. By relying on our categories and categorization, we can recognize that a situation is such that it calls for certain actions rather than others—e.g. to kiss a Sicilian padrino’s hand after a meeting; not to kiss your teacher after class.

Categorization can in general be thought as two-step process. For some item and some set of categories, the similarity of the item to each category is firstly computed. Then, these similarity ratings are transformed to determine the category to which the item belongs. “In general, a model of categorization specifies three things: (1) the content and format of the internal categorical knowledge representation, (2) the process of matching a to-be-classified stimulus to that knowledge, and (3) a process of selecting a category based on the results of the matching process” (Kruschke 2008a, p. 269).

The focus here is (1): what is the format, or formats, of the categorical knowledge representations underlying social norm compliance? I begin to tackle this question by considering exemplars.

1.1 Exemplars

Exemplars are bodies of knowledge about individual members of a category (Medin and Schaffer 1978). The wedding of Don Vito's daughter is an exemplar of the category "Sicilian wedding." Exemplars correspond to particular, actually experienced instances which we recall when we need to classify a novel item. Thus if we rely on exemplars to categorize Don Vito as a Sicilian "padrino," we retrieve the features of particular people we have encountered in the same type of situation, and compute for each person his similarity to Don Vito. I now consider how we come to acquire and use exemplars when we comply with a social norm by presenting Sripada and Stich's (2006) suggestion that norms may be stored as exemplars.

1.1.1 Sripada & Stich on Social Exemplars

Sripada and Stich (2006) suggest that social knowledge may be stored as exemplars. Norm compliance would depend on the representations of particular cases of norm abidance and norm breaking behaviour. These representations would contain contextual information about particular people behaving in situations at a given place and time.

On this account, when we face a new social situation—say the wedding of Don Vito Corleone's daughter—we judge which behaviour is (in)appropriate by retrieving stored wedding-exemplars—say your schoolmate's wedding in Vegas, Diana and Prince Charles's wedding, your Italian cousin's wedding, and so forth—and by evaluating their similarity to the current instance. If the current situation is mostly similar to a stored exemplar where behaviour of type *A* is inappropriate, then the current instance of *A* is likely to be judged to be inappropriate in the situation at hand.

People may search exhaustively through all their stored relevant exemplars and compare each of them to the behaviour to be evaluated. There may be no constraint on the exemplars-space to be searched: People would search and evaluate *all* of their stored exemplars in judging whether certain behaviour is appropriate in the current social situation. More plausibly, the set of exemplars taken into consideration may be constrained. One's cognitive and emotional history may prime a certain subset of stored exemplars which are used to categorize a new social situation and to generate judgements about which behaviour is appropriate. In fact we tend to recall more easily the first or the last few exemplars of a category we have encountered; emotionally charged exemplars are likely to be recalled more often, vividly and with more details than emotionally-neutral exemplars. As Sripada and Stich (2006, sec. 5.3) surmise, people may make different judgements about the same type of situation on different occasions in function of the subset of their stored exemplars primed by current circumstances and their cognitive and emotional history.

Sripada and Stich's (2006) acknowledge that their proposal is not backed by telling evidence. The type of argument in support of their suggestion is similar to the one sketched by Stich (1993). It has the form of an inference to the best explanation: Were social knowledge concerning norms stored as exemplars *instead of* tacitly known rules, some facts about social normativity would be plausibly explained. Stich (1993) indicates a number of explanatory payoffs that an exemplars-based account of social knowledge would have.

If social representations consist of clusters of stored exemplars, then the fact that some instances of social situations are easier to categorize and easier to recall would be easily explained because exemplar-based categorization is sensitive to situational factors that may prime one or another stored exemplar. Such sensitivity would also explain much of the variability of normative judgements concerning appropriate behaviour in a given situation. The same type of behaviour in the same type of situation can be judged differently in function of the subset of exemplars more vivid in memory and easier to retrieve. Finally, Stich (1993) emphasizes the pedagogical importance of myths, parables and fables. If the preferred format in which our social representations are stored is that of exemplars, then social and moral knowledge cast in the form of rules may be ineffective since social representations in this format would not be easy to build and use. Fables, stories and myths, instead, would be particularly effective, since they would furnish our memory with a rich stock of social and moral exemplars which can be more readily used to judge and act appropriately in new social situations.

Sripada and Stich (2006) leave open the (likely) possibility that people use a variety of representational formats to store and recall knowledge important to categorize new social situations and comply with social norms. Social prototypes, exemplars, theories and narratives might be activated in function of different contexts. For example, Sripada and Stich speculate, exemplar-based processes might be primarily involved for categorization of socially appropriate behaviour "in the context of day-to-day norm-related cognition, especially when such judgement are made rapidly and 'on the fly'" (Sripada and Stich 2006, p. 293).

1.2 Prototypes

Prototypes are knowledge summaries extracted from information about the individual members of a category (Rosch 1978). Such summaries could be bodies of statistical knowledge about the features that are typical or diagnostic of events, individuals, and objects in a category. They describe a central tendency that can be expressed as an average of the category. This average need not correspond to any particular actually experienced instance.

If we use prototypes when we categorize Don Vito as a Sicilian “padrino,” we need not retrieve the features of any particular Sicilian godfather we have ever met. We may retrieve instead the standard, average, typical Sicilian godfather and compute his similarity to Don Vito. Plausibly, this prototype is the result of an average from the sample of all the Sicilian godfathers we have encountered. How could we come to acquire and use prototypes when we comply with a social norm? I answer this question by considering Paul Churchland’s account of social prototypes.

1.2.1 Churchland on Social Prototypes

Paul Churchland (1995, Ch. 6; 1998) defends the idea that knowledge of our social and moral world is represented as a family of prototypes embodied in the specific configurations of the many synaptic connections between neuronal layers. Chapter 5 will expand on Churchland on moral thought, for the moment let’s focus on his argument concerning the representation of social and moral knowledge. On his account, social representations are stored as clusters of prototypes that carry information about especially typical examples of actions that are required or prohibited by the relevant social norm. Churchland makes two claims: first, social and moral knowledge is stored in the nervous system as learned prototypes; second, prototypes are vectors (i.e. order sets of numerical values) that describe the structure of connections between neurons.

One example used by Churchland to explain and support these claims is EMPATH: an unsupervised neural network that can recognize emotions from human facial expressions (Cottrell and Metcalfe 1991). After a training session where the network was presented with twenty pictures of faces each displaying eight different emotions (160 pictures of faces in all), EMPATH could develop prototypical patterns of activation associated to facial expressions. Drawing on its prototype-style body of knowledge, EMPATH could achieve near perfect rate of successful discrimination between male and female faces; it could also successfully identify five out of the eight types of emotional expression. EMPATH had some limitations as well. In particular, its capacity for generalization to new faces was poor. Churchland maintains that, all in all, EMPATH provides an "existence proof" that nervous systems can learn to generate behaviourally appropriate outputs in social contexts by using knowledge stored in a "library of social prototypes" (Churchland 1995, p. 127).

Churchland's connectionist account of social and moral knowledge has a number of interesting consequences that can help us to understand to what extent our knowledge of social norms is represented in prototypes. First, if we store social representations as prototypes, then social learning involves extensive training with numerous, distinct situations that display a variety of social features. We acquire social prototypes by repeated exposure to and practice with various examples of a given category. The training leading to the acquisition of prototypes constitutes a learning history which causes internal changes in certain populations of neurons.

Learning histories in social environments can in fact cause changes in synaptic connectivity between neurons. These changes can consist in the growth of new synapses on existing neurons or in chemical alterations in existing synapses. For an artificial neural network like EMPATH, the learning history causes particular sets of weights between processing units to become more stable. Each set of weights can be described with an ordered set of numerical values, that is, as a vector. The values in each vector correspond to variable social features; they may correspond, for example, to variable features of an action or to variable features of one's facial expression like mouth width, eyebrow position, eye gaze direction, and so on. When, as a consequence of one's learning history in a type of social situations, the features common to those situations become strongly associated courtesy of the formation of mutually excitatory links across some units of the network. The connection weights between these units tend to be stronger and encode specific, more stable values. The weight structure of the net thus becomes a background condition that enables the reliable detection of prototypical social features across situations of that type. Such prototypical social features can be used to categorize and to know what to expect from future social situations.

By categorizing a given social situation as a "tutorial class," for example, we know what to expect from others and what others expect from us. If you mistakenly categorize a tutorial class as a "punk concert" and start to scream and to jump up and down, the people around may stare at you baffled.

Second, social prototypes need not correspond to any particular example of a category. When you categorize a social situation by relying on prototypes, the specific examples from which the prototypes were extracted need not to be internally represented. During the processes of abstraction and storing over your learning history, much of the information concerning specific instances is discarded. A "Sicilian wedding party" prototype might include features such as pictures of Saints or statues, a lavish feast with one large family, a big meal outdoor, traditional music and dancing. It need not contain contextual information about particular people, places or times.

Third, the processes supporting the encoding of a social prototype are not specific to morality or social normativity. The difference between different kinds of prototypes depends on the type of set of training instances taken as input by the learning network. In the case of social normativity, this set comprises social features, whereas in other cases the training instances are purely physical features that may not concern any aspect of social behaviour. This suggests that there is no specific function computed by the activity of some neural circuit dedicated to the acquisition and storage of social knowledge.

Fourth, if social normative knowledge is encoded in prototypes, then fables, cartoons and parental example play an important causal role in building a stored library of "learned prototypes." Fables, myths, cartoons and daily examples of appropriate social behaviour would be our main sources of moral and social prototypes. Social education would strongly rely on such sources because our cognitive system would be best suited to learn and use information in the form of prototypes.

Finally, Churchland's account suggests that socially virtuous people are those who possess a bundle of perceptual and behavioural skills. Such skills depend upon the acquisition of a rich library of diverse social prototypes which can be used to comprehend one's and other's social situation. By relying on their prototypes, virtuous people can see the same social situation from different angles, and correctly evaluate the appropriateness of different ways to interact with other people. Virtuous people, that is, can swiftly and successfully navigate the high-dimensional social space by recognizing their and other people's position in it. Social misbehavior instead would primarily depend on a socially deprived or highly biased learning history. If the sample of training examples one is exposed to during learning is very small or highly skewed, then the resulting library of social prototypes will be excessively scant and unvarying. The lack of a rich and diverse library of social prototypes may cause a kind of perceptual failure which consists in an incapacity to appreciate the full range of dimensions and structure of the social domain. One likely result of this perceptual handicap is the failure to comply swiftly and reliably with social norms.

1.3 Scripts

Scripts (or schemata) are rich bodies of causal, functional, and nomological knowledge about categories of complex situations. Scripts specify sequences of events and actions that characterize the typical structure of well-known situations such as "a lecture," "a birthday party," or "a wedding reception" (Schank and Abelson 1977). For example, a script of some wedding reception may consist of a rundown of a typical sequence of events like toasting, cheering and dancing; it may comprise information about cakes, dresses, music, guests, family and friends. Scripts capture background knowledge about a given type of situation, enable us to make sense of it and behave appropriately. If we use scripts when we make sense of Don

Vito's behaviour during his daughter's wedding, then we do not compute a similarity rating between the current situation and past situations we have encountered, as we would do if we used exemplars or prototypes. The use of scripts relies on pattern completion functions. The activation of a subset of a stored pattern of events corresponding to a "Sicilian wedding reception" script triggers the filling in or completion of the remaining portion of the pattern. More on this in a moment.

Scripts appear to be knowledge structures more complex and more computationally expensive than exemplars and prototypes. Apart from early script-based approaches to categorization and knowledge representation (Minsky 1974; Schank and Abelson 1977), more recent theories of categorization based on scripts have had limited formalization, partly because of the difficulty to formally specify all the relevant details of a complex knowledge structure (Kruschke 2008a).

Although we may doubt that our cognitive system employs such rich and computationally heavy bodies of knowledge, we should consider that scripts or schemata need not be explicitly stored neither need they cover all possible contingencies of a situation. Scripts can be modelled as knowledge structures emergent from the activity of a neural network that responds to the presence or absence of relevant microfeatures (Clark 1989, Ch. 5.4). From this perspective, "[t]here is no representational object which is a schema. Rather, schemata emerge at the moment they are needed from the interaction of large numbers of much simpler elements all working in concert with one another" (McClelland, Rumelhart, and the PDP Research Group 1986, p. 20).

I mentioned that the type of processes underlying the learning and application of scripts (or schemata) in neural network involve pattern completion functions. The input pattern in this case spans rundowns of sequences of events instead of simple examples. Such sequences of events consist of ordered patterns of microfeatures. Given repeated exposure to complex patterns underlying a given type of situation, neural networks can learn a schema by settling on certain connectivity weights. The connectivity weights learned by the network are such that they respect as far as possible the possible relationships and constraints associated to the ordered microfeatures corresponding to the sequence of events. Once a script has been learned, the presence or absence of some microfeatures activates a subset of a "known" pattern in the network. Such activation can be sufficient for the network to fill in or complete the remaining portion of the pattern in a way maximally coherent to its connectivity weight structure. Thus the network settles on a particular activation pattern from which the properties of the script emerge. Let us now focus on how clusters of social representations can be understood as scripts or schemata.

1.3.1 Bicchieri on Scripts and Social Norms

Bicchieri (2006, p. 96) argues that "*social norms are embedded into scripts*." She understands schemata and scripts in terms of "theories of the way social situations and people work" (p. 81). Such theories enable us to navigate our social world because they support inductive inferences and predictions about people's behaviour.

Bicchieri distinguishes categorization from script activation. Categories are knowledge structures that contain information about instances of the items of a class (e.g. the class of "waitresses"). Categorization, for Bicchieri, activates scripts which are knowledge structures that contain information about the attributes and relationships among categories (e.g. the script "dinner at a restaurant in Japan").

Knowledge directly relevant to social norm compliance would be stored in scripts or schemata that “contain social roles and expected sequences of behaviours that help us to behave appropriately (and know what to expect) in specific settings” (p. 82). Bicchieri emphasizes that scripts and schemata need not be explicitly stored and need not be accessible to consciousness.

Like Sripada and Stich (2006), Bicchieri does not provide direct support for her claim by drawing on some particular experimental finding. The form of her argument is an inference to the best explanation with the following form: There are a number of facts related to social normativity. If norms are embedded into scripts, then many facts related to social normativity would be explained. Therefore, it is plausible that social norms are embedded into scripts.

It is noteworthy that while Sripada and Stich point to linguaform rules as *prima facie* rival hypothesis to exemplar-based social representations, Bicchieri does not point to any relevant alternative hypothesis to scripts. But the validity of inference to the best explanation is sensitive to the pool of explanations under consideration. The introduction of some relevant alternative explanation can invalidate the validity of a plausible inference to the best explanation even when the empirical evidence has remained unchanged.

This said, what are the *explananda* that scripts-based norm compliance would explain? Bicchieri singles out at least three *explananda*. The first fact recalled by Bicchieri is the difficulty in defining “general principles of fairness, or justice” (Ibid., p. 95). If we reason through schemata and scripts, then it is plausible that the meaning of e.g. “fair division” is understood by means of sequences of events in familiar situations involving certain divisions of a good. What is taken to be “fair” would depend on knowledge about clusters of categories of particular people, events, and objects. Given the variety of the categories activated in a social exchange that we describe as "fair," and given that the particular members of those categories may be very different from each other along many dimensions, there may be no context-invariant features captured by general principles of fairness.

For Bicchieri, two other *explananda* would be explained if social knowledge is mainly embedded into scripts: what grounds the projectibility of certain behavioural patterns and what it is that confers legitimacy to other people's expectations in certain social interactions (Ibid. 95-6). Consider Don Vito receiving pleas for favors in his private office during his daughter's wedding. Why do his guests and his family perceive his behaviour as appropriate and legitimate? Why is that behavioural pattern taken to be projectible to future situations? We can answer both questions by appealing to scripts and considering that social interactions embedded in scripts tend to be *regarded as* "natural kinds"—classes that represent some real distinction in nature and that support inductive inferences.

If script-based social interactions are regarded as natural kinds, then scripts would ground people's expectations concerning social situations. We would believe and expect certain things in a situation in function of the script we have activated. Since social norms, according to Bicchieri, are sets of mutual expectations, when particular expectations come to be prompted by the activation of a script, the behavioural regularity underlain by those expectations is automatically projected: "It's part of the wedding" explains Tom Hagen.

The attribution of legitimacy to the expectations underlying that behaviour would also be explained by our propensity to regard scripted social interactions as natural kinds. The existence of a script that represents knowledge about a type of situation is the source of legitimacy. If receiving quests for favors during your daughter's wedding is embedded into a script, then the guests will believe it legitimate to ask for favors and to obtain them; and they will be angry if their expectations are frustrated.

2. Exemplars, Prototypes or Scripts. What Difference Does It Make?

Do people store social representations in a single format? To address this issue I review evidence from the cognitive neuroscience of categorization and category learning. Much of the results I present involve non-social, non-moral information. This is for two reasons. On the one hand "the empirical study of the representational format of norms has barely begun" (Sripada and Stich 2006, p. 293). On the other hand, the stimuli used in experimental tasks of categorization and category learning often consist of artificial objects characterized only by their perceptual properties. This is mainly to control for the effects that knowledge possessed by subjects about a domain may have on learning and categorization.

2.1 Category-Learning and Categorization Tasks

Imagine that you barely know Don Vito Corleone, yet you happen to be at his daughter's wedding party and you must judge whether that particular circumstance is an instance of the category "Sicilian wedding."

As a consequence of having participated to many weddings, you may have abstracted from particular instances a prototypical general tendency of various wedding categories—for example, based on the types of religious signs, music, and food you have encountered in each wedding you have participated to. You note that the current situation is most similar to the prototype of a "Sicilian wedding" rather than "Jewish wedding" or "Polish wedding." On this basis, you categorize the current situation as an instance of "Sicilian wedding." This is, in a nutshell, the sequence of processes involved in categorization based on prototypes.

Categorization based on exemplars involves different processes. Because you have participated to many weddings, you may have stored many wedding exemplars in your long term memory. You notice that the current situation is most similar to the stored exemplar of "Angelica's Sicilian wedding." Drawing on such a similarity, you conclude that the wedding of Don Vito's daughter is an instance of "Sicilian wedding" and thereby you can make sense of the situation and understand which behaviour is appropriate.

In relation to scripts, Bicchieri (2006) argues that people interpret and categorize a given context in function of the situational cues, or microfeatures, that spark their attention. The processes underlying script-based judgement rely on *spreading activation* and pattern-completion. The activation of the representation of a certain complex situation spreads to representations of situations related to it. Social categorization activates scripts that enable us to understand social situations, to predict others' behaviour and to respond appropriately to their actions. Scripts—recall—are theories that represent generic knowledge about well-known classes of situations. According to this theory-based approach, you judge whether the wedding reception of Don Vito's daughter belongs to the category "Sicilian wedding" by determining whether the features of that instance are best accounted by the theory underlying that category. Let’s now leave real-world, intuitive cases, and enter the lab.

In typical categorization and category-learning tasks, experimental subjects are required to learn and use some category. The task is generally in three phases. In a learning phase, the subjects are presented a number of items and are informed under which category each item falls. During this category-learning phase, the task of the subjects is to acquire some body of categorical knowledge from encountering some members of the extension of the relevant category. In a test phase, the subjects are presented both with items they had already encountered during the learning phase and with new ones. This is, strictly speaking, the categorization task which consists in judging whether certain items belong to a given category or whether some classes are included in a given category. Finally, a recognition memory task may follow. The subjects are asked to discriminate between “old,” already encountered items, and new ones. What may this type of task tell us about prototypes exemplars and scripts?

2.2 Exemplars

If we use exemplars instead of prototypes, then at least four empirical predictions follow with respect to people’s performance and its underlying mechanisms in category learning and categorization tasks.

First, the learnability of a category measured in terms of the time needed to learn that an item belongs to the category will not depend on the typicality of the item. It will depend on its similarity to known members of the category. In comparison to a typical item that is not similar to previously encountered category members, we would learn more quickly that a less typical item belongs to a certain category, if this item is similar to previously encountered category members. It would be quicker to learn to categorize a woman pastor as a pastor than a man pastor because your sister is a pastor (on this effect see e.g. Medin and Schaffer 1978).

Second, the same would apply to categorization performance measured in terms of reaction time and accuracy. The time employed by people to categorize an item and their accuracy would not depend on the similarity of the item to the prototype of the category. Less typical items would be categorized more easily and accurately if they are similar to already stored exemplars.

Third, during a recognition task, old items would have an advantage over equally typical but new items. It would be easier to categorize your friend Don Vito as a Sicilian godfather than an unknown Sicilian godfather that is an equally typical Sicilian godfather (on this type of old-item advantage see e.g. Nosofsky 1992).

These effects suggest a fourth neuropsychological prediction. The same representations that enter the process of categorization would also be involved in recognition memory tasks. If categorization is exemplar-based and relies on the same representations involved in recognition tasks, then amnesic patients will exhibit abnormal performance in categorization. Let us expand on this type of prediction by presenting a famous case study.

Amnesic patients are impaired both in the ability to store new representations in declarative memory and in the ability to verbalize knowledge of exemplars already encountered. They typically display severe injuries in the medial temporal lobes in both hemispheres. Squire and Knowlton (1995) tested the hypothesis that no category learning should take place without the capacity to store exemplars (see also

Knowlton and Squire 1993). They examined the performance of a severely anterograde and retrograde amnesic patient, E.P., in a learning and categorization task. E.P. couldn't recognize previously encountered objects, which suggests that he couldn't acquire and store representations of new objects. Squire and Knowlton found that in spite of such impairment E.P. could perform normally in a dot-distortion category task. In this task subjects are typically presented with patterns of nine dots generated by randomly distorting one of a number of prototype-patterns which define different categories (Posner and Keele 1968). In the test phase subjects are asked to classify both new patterns and patterns they had already encountered. Squire and Knowlton's (1995) subject exhibited zero ability to recognize whether a given item was a new or an old, already encountered, exemplar. However, E.P. performed normally on the categorization task: E.P.'s categorization judgements were a function of the typicality of the target pattern.

E.P.'s performance is hard to explain by appealing to knowledge stored in declarative long term memory since the patient had no declarative memory abilities whatsoever. The patient must have used a categorization procedure different from an exemplar-based procedure. During training, E.P. could have learned a prototype of the category of dot patterns and retrieved this representation to categorize new patterns. Squire and Knowlton conclude: "These findings demonstrate that the ability to classify novel items, after experience with other items in the same category, is a separate and parallel memory function of the brain, independent of the limbic and diencephalic structures essential for remembering individual stimulus items (declarative memory)" (Squire and Knowlton 1995, p. 12470).

This conclusion is coherent with the results found by Kolodny (1994). In this study, amnesic subjects were tested in the dot-pattern task, but also in a task designed to elicit exemplar-based processes. In this latter task, paintings of three Italian Renaissance artists were presented to the subjects who were required to learn which paintings were made by the same artist. The exemplars of each category lack obvious stylistic relations that could facilitate the acquisition of a prototype for each artist. Hence, it is plausible that such categories are learned courtesy of explicit memorization by storing exemplars after extensive experience. Amnesics' performance in both learning and categorizing paintings was at chance. Unsurprisingly, they also performed poorly in the memory recognition task.

From these behavioural and neuropsychological results the following predictions might be extracted about social cognition. If people store social knowledge relevant to norm compliance in a single exemplar-based format, then judging which social context one is facing and whether an action is appropriate in that context will engage long term declarative memory. If structures supporting long term declarative memory are impaired, as in amnesic patients, we may expect inappropriate social behaviours also in situations already encountered.

2.3 Prototypes

Let's now consider prototypes. If we use prototypes instead of exemplars, then at least four empirical predictions follow with respect to people's performance and its underlying mechanisms in category learning and categorization tasks.

First, the learnability of a category will not depend on whether the item members are similar to some already encountered items. It will depend on their similarity to the typical member of the category. We would learn more easily and quickly to classify a typical member of a category that has not been encountered during training than other non-typical members seen during training (on this effect see Posner and Keele 1968).

Second, categorization performance would depend on the similarity of the item to the prototype of the category. Most typical items would be categorized more accurately than other typical items even if they have not been already seen (Smith 2002). The prototypical central tendency shared by the items we have encountered may give rise to a kind of "perceptual fluency." After some experience, people may experience a sensation of fluency in categorizing exemplars that are most similar to a prototype.

Third, during a recognition task, old items would not have an advantage over equally typical but new items. The recognition of an item would depend more on its typicality than on the fact that it has been previously encountered. Because perceptual fluency may be based on perceptual inaccessible processes, people often cannot do any better than recalling general features defining a prototypical tendency to justify the basis of their categorizations and recollections.

Fourth, prototypes would engage declarative memory storage less than exemplars, as they need not contain any contextual information. A prototype might be abstracted and used by relying on knowledge that cannot be easily verbalizable. This would explain why amnesic patients are successful in dot-pattern categorization tasks but not in recognition which requires explicit, declarative memory. Retrieving representations with contextual associations requires an intact medial-temporal-diencephalic system (Smith 2008).

In light of these considerations, if people store social knowledge relevant to norm compliance in a single, prototype-based format, then they will perform poorly in unstructured situations such as the painting task where exemplars don't share any obvious feature. If the social situations that one encounters don't share any apparent pattern, it may be difficult to classify them by using a prototype. If amnesic patients can acquire and use prototypes and their social knowledge is stored as prototypes then they are *not* likely to behave inappropriately in typical social situations.

2.4 Scripts

A script-based account of social categorization has barely been investigated in cognitive neuroscience. This may be because scripts are complex knowledge structures of difficult computational formalization (Kruschke 2008a). Scripts (or schemata) contain information organized in large clusters that serve to generate inferences. Their activation is likely to depend on a number of mechanisms that support such functions as semantic knowledge, declarative and "implicit" memory, cognitive control, evaluation and information integration. It seems hard to isolate precise empirical predictions from the hypothesis that norms and social knowledge are embedded into scripts. Given the diversity of the cognitive functions that are likely to be involved in script-activation, it is probable that the prefrontal cortex (PFC) is essential for storing and using script-based social knowledge relevant to norm compliance.

Krueger et al. (2009) offer a framework to understand how complex knowledge structures akin to scripts and schemata are supported by brain activity in the PFC (see also Grafman 2002; Wood and Grafman 2003). They argue that "the [medial prefrontal cortex] mPFC represents '*e*vent simu*lators*' (elators) that give rise to social event knowledge via structural and temporal binding with regions in the posterior cerebral cortex and limbic structures" (Krueger et al. 2009, p. 103). Before examining Krueger and colleagues' proposal, it is worth repeating a theme that will be reiterated in Chapter 5: the precise computational architecture of the PFC is poorly understood. The mPFC comprises distinct, functionally diverse regions—the medial orbitofrontal cortex, ventromedial prefrontal cortex, dorsomedial prefrontal cortex—which have been found to be involved in many different social and non-social tasks whose solutions may require the computation of distinct functions (see e.g. Fuster 2008; Miller et al. 2002). We should be wary about claiming that the PFC is engaged in *particular* tasks and computes *particular* functions.

For Krueger and colleagues, elators are abstracted from experience with multiple exemplars of social situations. Given the complexity of such knowledge structures, it is likely that acquiring and using elators engage various mnemonic abilities. Information about a social situation might be first stored as an exemplar associated with a specific place and time. With repetition and experience, such information might be involving semantic memory which stores our knowledge of the world, and procedural memory which store "implicit" knowledge of skills like driving a car. It is not clear whether elators' formation relies on implicit prototypes, on exemplars or on both. In default of a detailed account, it remains difficult to assess what kind of evidence would count against the claim that "elators are abstracted from experience."

Krueger and colleagues define "abstractions" as "dynamic summary representations," which are also called "structured event complex" (Forbes and

Grafman 2010, p. 311; see also Wood and Grafman 2003). These abstractions, or structured event complexes, are set of events linked together to form a script or schema. They embody general knowledge about how situations unfold. More precisely, they can encode goal- or outcome-oriented set of events ordered sequentially around thematic activities such as “Checking in at the airport” or “Attending a lecture.” Goal-oriented knowledge, according to Krueger and colleagues, is about the likely actions that agent will take when they desire to accomplish a task or reach a certain aim. Outcome-oriented knowledge mainly concerns the likely affective response to goal attainment.

According to Krueger and colleagues, these types of knowledge structures guide our behaviour and perceptions by embodying information about social groups and norms. Knowledge about social norms and social groups would be localized in the left anterior ventromedial prefrontal cortex (VMPFC), related to outcome-oriented events. Forbes and Grafman (2010, pp. 312-3) claim that “the VMPFC stores structured event complexes specific to social norms and scripts.” The evidence they provide for this claim is from studies that point to the involvement of the VMPFC in stereotype-based judgment. In particular, compared to healthy subjects and patients with damage to the dorsolateral prefrontal cortex, patients with VMPFC damage show reduced levels of stereotyping when gender-bias is assessed through an implicit association task (Milne and Grafman 2001). This indicates that the VMPFC may be necessary to automatically retrieving some aspects of (implicit) social knowledge, but does not give us strong reason to believe that VMPFC is the circuit where scripts embedding social norms are stored. Milne and Grafman’s VMPFC patients displayed normal explicit knowledge of gender stereotypes, moreover the authors are careful in pointing out that their evidence is insufficient to distinguishing whether VMPFC patients' deficit is "specific to social knowledge (versus other forms of stimulus-response compatibility)" (Milne and Grafman 2001, p. 5). Finally, it is likely that the cortical representation of scripts embedding social norms is distributed across several neural networks comprising the amygdala and the orbitofrontal cortex besides the VMPFC (e.g. Casebeer and Churchland 2003).

VMPFC might be necessary to respond smoothly to some contextual social cues and prime certain structured event complexes. The presence of particular people in some types of situations might prime scripts associated to those cues. Impairment in the capacity to automatically retrieve certain scripts—as in the case of VMPFC patients who seem to be insensitive to cues leading to implicit gender bias—may lead to inappropriate behaviour. Patients with VMPFC lesions often display a lack of compliance to social norms (see e.g. Dimitrov et al. 1999). "It may be—as Milne and Grafman (2001, p. 6) conclude—that a contributing factor to that social conduct impairment is the inability of those patients to automatically and rapidly associate differing aspects of social knowledge—a form of social agnosia."

In light of these considerations, if people store social knowledge relevant to norm compliance in a single, script-based format supported by the activity of VMPFC, then we can draw at least four predictions. First, patients with ventromedial damage will show deficits in storing and retrieving social information that supports social norm compliance. Second, because of damage in the VMPFC, subjects will display poor performance in social tasks that require the activations of social knowledge structures that are goal- or outcome-oriented and temporally ordered. Third, they won't be able to learn new social norms embedded in social scripts.

Fourth, they will have troubles in being sensitive to cues in situations that call for appropriate behaviour.

3. Representational Pluralism from a Probabilistic Approach

Research on categorization and human category learning has entered a "second generation" (Ashby and Maddox 2011). During the first generation, from the 1990s to the early 2000s, research in cognitive neuroscience addressed the question of whether there are multiple systems for categorization and category learning. Many researchers are now persuaded that there is telling evidence for multiple category-learning systems (Smith and Grossman 2008, for a review). As a result, according to Ashby and Maddox (2011), "second-generation questions" have begun to be tackled. These questions start with the assumption that humans store and use bodies of knowledge in multiple formats for categorization and category-learning. This chapter concludes by making the same assumption. I argue for an approach that can be fruitful to explore how inductive inference underlying social categorization and norm compliance can draw on bodies of knowledge that can take a plurality of formats. I start by elaborating on the nature of the problems of categorization and category learning.

Categorization and category learning are problems that require uncertain conjecture from partial, noisy and ambiguous information. They can be understood as inductive inferences that we draw about the organizing structure of a dataset. Inductive inferences can be understood as computations on uncertain sensory input data. In the social case, categorization and category learning can be understood as computations on uncertain sensory input data that lead to the discovery of relationships between agents, objects and events in our social landscape. The organizing structure of a dataset is provided by such relationships. These relationships can correspond to structured, non-arbitrary classes of agents, objects and events, or to structured classes of classes. They can correspond, that is, to social categories or to systems of categories. Information about these relationships allows us to build complex systems of knowledge about our social world and its underlying regularities. From this perspective, one of the deepest challenges in understanding social categorization and social category learning as types of inductive problems is this: How can we build complex systems of social knowledge from the sparse data yielded by our sensory systems?

This challenge can be addressed with the probabilistic approach we have already encountered when I explained the Bayesian mechanism that might underlie the acquisition of social representations. By focusing on the notion of *structural form* I now explain how this approach emphasizes the importance of representational diversity (Griffiths et al. 2010; Tenenbaum et al. 2011). I suggest that a probabilistic approach is particularly fruitful for evaluating proposals about the different forms social knowledge can take across different domains and tasks. Consider once again the case of the wedding reception of Don Vito's daughter.

This situation generates a stream of sensory input data. Given the data set of your sensory inputs in that situation, you need to infer what type of situation you are facing so that you can understand what types of actions are appropriate there. The problem is that "any finite set of data is consistent with an infinite number of inductive hypotheses" (Holyoak 2008, p. 10637). Different hypotheses about the situation at hand are available to you—you can interpret it as a barbeque party in fancy dress, as a Jewish wedding, or as a Sicilian wedding reception. The dynamics of our navigation in that situation will depend on the hypothesis we select.

The selection of one hypothesis rather than another might be carried out by a Bayesian mechanism. Chapter 1 made the suggestion that the acquisition of social representation might depend on a Bayesian mechanism. This mechanism would be a statistical inference engine that integrates abstract knowledge encoded in a probabilistic generative model with data from different sensory sources. The abstract knowledge supporting the inferences drawn by such a Bayesian machine can take multiple forms. The Bayesian machinery, that is, is not committed to process representations in a particular format. It works on probability distributions over observable data which can take any form. Before articulating this last point, let me clarify the role of abstract knowledge in the probabilistic approach I am describing.

The body of knowledge that guides social categorization, social category learning and social norm compliance needs not be specific to the particular situation at hand. It concerns whole classes of situations over which experience gained in a particular case can be used to make predictions and take appropriate actions. This body of knowledge captures the essential structural form of situations giving rise to the agent's sensory input data. More precisely, knowledge about the essential structure of the situations we encounter is embodied in a constrained space of hypotheses that could explain the sensory data generated by a given situation. Each hypothesis comes with a certain probability distribution. The probability distribution specifies the agent's degree of belief in a specific hypothesis about a structural form underlying a situation prior to the observation of sensory data. By combining prior hypotheses and sensory data in a Bayesian fashion, agents can come to identify the hypothesis that account best for the data: the hypothesis that has highest probability conditional on the data. Identifying the structure underlying a situation provides us with significant constraints on our inductive inferences. Granted that bodies of abstract knowledge encoded in a probabilistic generative model constrain and guide social categorization and category learning, what does it mean that they 'capture the essential structural form of a situation'? And what is the form of these bodies of knowledge?

In Chapter 1 I mentioned that states in the environment stand in causal relationships and that these causal relationships can be referred to as *structure*. Different relationships between states, different structures, can be depicted by means of graphical models. More generally, the structure of a situation consists in its underlying regularities. These regularities need not be causal. They can be conceptual, temporal, or spatial, for example. Hence, in a general sense, the structure of a situation needs not be causal. Different structures—either causal or non-causal—underlying a situation can be depicted by means of graphical models, for example: partitions, chains, trees, grids and cylinders. So, to say that a body of knowledge captures the essential structure of a situation is to say that they contain information about the causal or non-causal relationships between the individuals, events and objects that constitute that situation. Such relationships can be represented as a tree, for example, with nodes and edges constituting a particular structural form.

Griffiths et al (2010, p. 358) claim that "connectionism makes strong pre-commitments about the nature of people's representations and inductive biases based on a certain view of neural mechanisms and development: representations are graded, continuous vector spaces, lacking explicit structure, and are shaped almost exclusively by experience through gradual error-driven learning algorithms." In a *purely* bottom-up, connectionist approach, that is, background knowledge is encoded in continuous vector spaces which lack explicit structure. These vector spaces describe the connectivity weight structure of the network which can embody social representations typically in the form of prototypes or schemata. At best the connectivity structure weight of a network can only approximate in an implicit fashion representational forms like trees or hierarchies that people appear to know and use explicitly (Griffiths et al. 2010, pp. 359-360; Gopnik et al. 2010).

In comparison to a purely bottom-up connectionist approach, the probabilistic approach makes no *a priori* assumptions about the form of social representations. Probabilistic models are apt to explore a larger space of representational possibilities. Representations in different formats can in fact be needed for different types of inferences underlying different cognitive functions. Kemp and Tenenbaum (2008), for example, showed how qualitatively different representations can explain human inferences in many different real-world domains. Inductive inference about different real-world domains seems to be best explained by appealing to representations with different structural forms (Kemp and Tenenbaum 2009). In a probabilistic approach, the fact that background knowledge is encoded in probabilistic generative models does not mean that the hypotheses constituting the background knowledge must be in a single particular representational form. The format that hypotheses and background knowledge can take span from weights in a neural network to structured symbolic representations. Now, how should we assess the claim that a probabilistic approach is fruitful to understand what is the representational format of the background knowledge that supports the inductive inferences underlying norm compliance?

By operating on a broad range of candidate representational formats, probabilistic models can generate interesting empirical research also in the social and moral domain. We are interested in understanding the representational format of social norms; we are interested in identifying under what circumstances representations in a given format support social norm compliance. One way to do this is shown by Kemp and Tenenbaum (2008): A probabilistic model may be defined and social representations in a particular format specified within the model. When the probabilistic model does not fit behavioural data concerning, for example, social categorization and category learning, we may use a qualitatively different representation while retaining the explanatory framework of Bayesian computation. This will enable us to identify which representational format best explains behavioural performance in the social domain. We can thus evaluate different proposals within the same type of probabilistic explanatory framework.

The structured representations that might be used in probabilistic computations in the social domain need not be explicitly encoded in the brain. There is a growing wealth of research on how the brain might maintain a generative model of the environment, and how neurons might encode probability distributions and combine those distributions according to close approximations to Bayes' rule (e.g. Berkes et al. 2011; Ma et al. 2006). Yet we are far from understanding how exactly representations in multiple formats supporting Bayesian inference are encoded in neural circuits. This "is arguably the greatest computational challenge in cognitive neuroscience more generally—our modern mind-body problem" (Tenenbaum et al. 2011, p. 128).

Conclusion

This chapter has distinguished three specific options about the format in which social representations relevant to norm compliance may be stored. I have discussed exemplars, prototypes and scripts, and related each option to the social and moral domain. For each option, empirical consequences have been drawn. After having noted that there seems to be telling evidence that we use representations in multiple formats, I have presented in broad strokes a probabilistic approach to cognition that allows for representations in multiple formats. I have argued that this approach, in comparison to a purely bottom-up, connectionist one, is probably more fruitful to understand what is the representational format of the background knowledge that supports the inductive inferences underlying norm compliance.

CHAPTER 4.

Moral Judgement for Bayesian Brains

This chapter argues for two claims. First, some central aspects of the psychological mechanism of moral judgement can be described within the RL-Bayesian neurocomputational framework laid out in Chapter 1. Second, such a neurocomputational description of moral judgement can shed new light on puzzling findings about specific patterns of moral judgement.

The chapter builds on the account of the Bayesian brain put forward in Chapters 1 and 3 and on the notion of social (neural) representation characterized in Chapters 1 and 2. In spite of my reference to brains and neural representations, the discussion here will be at a more abstract level most of the time.

There are three sections in the chapter. The first section distinguishes between two broad senses of 'judgement.' The second section identifies three neurocomputational ingredients, which can be used to describe aspects of the psychological mechanism of moral judgement. Such ingredients are: the norm prior, the likelihood of moral judgement and the continuous updating of norms courtesy of Bayesian inference. The last section argues for new ways of understanding traditionally controversial findings concerning psychopaths' moral judgement and the ontogenesis of moral judgement.

A neurocomputational perspective on moral judgement promises to bear explanatory fruit because it forces us to move beyond either-or dichotomies, which have shaped and in some cases limited debates in the psychology of moral judgement. I have in mind such dichotomies as: emotion versus cognition, learned versus innate, rule-governed versus rule-free, moral norms versus conventions.

1. Judgement as a State and Judgement as a Process

‘Judgement’ is an ambiguous term. Different senses are hardly made explicit in discussions of moral judgement. I now distinguish between two general ways of understanding the term, which will be helpful to avoid confusion in the account articulated in the remainder of the chapter.

One way to understand ‘moral judgement’ is in terms of a mental *state*. In this sense, moral judgement can refer to either representational or non-representational mental states. Accordingly, ‘judgement’ can refer to mental states that are not necessarily representational and that can be expressed by sentences or utterances. In a narrow sense, ‘judgement’ refers to representational mental states. Beliefs are the paradigmatic example of such mental states. Beliefs are generally considered to be mental states that represent something to be the case. Whenever we take something to be the case or take it as true, we believe that something. So, moral judgement may refer to some moral belief we have. That a mental state is representational does not entail that we must be aware of that mental state. Whatever the representational status of moral judgement, a separate issue is whether we are aware of the moral judgement we entertain or not. ‘Moral judgement’ need not refer to states of the mind, representational or not, that involve active reflection or awareness of anything specific.

The second way to understand ‘moral judgement’ is in terms of a mental *process*. As a process, ‘moral judgement’ can refer to *deliberation* (or practical reasoning), which is not necessarily a conscious process. Deliberation (or practical reasoning), as understood here, is the process that enables agents to answer the question of what one ought to do. Courtesy of moral judgement (*viz*. deliberation), agents come to entertain such mental states as beliefs and attitudes expressing that one ought to behave in some way rather than another under certain types of circumstances.

Here I am concerned both with the *process* enabling agents to resolve what one ought to do under a certain type of circumstance, and with 'moral judgement' as a mental *state* expressible with a sentence or utterance. In neither of these senses, 'judgement' refers necessarily to an introspectively accessible or conscious mental state or process. As I go along articulating my proposal, I shall make clear which sense is relevant to my argument.

Representations are an essential part of the account of norm compliance I put forward in Chapter 1. Chapter 2 argues that we should explain norm compliance by appealing to neural representations. It is worth noting, however, that the centrality of the notion of representation in my account does not mean that I maintain that in our cognitive system *all* signals, which can affect our social/moral behaviour, must be representational. As I build on those two chapters here, I shall sometimes characterize moral judgement in terms of representations. This does not entail that the states of the mind expressed by moral utterances are necessarily beliefs or that the processes underlying moral deliberation are necessarily "cognitive" as opposed to "emotional." In fact, by embracing a neurocomputational perspective—let me emphasize it—one of the burdens of this chapter is to show some ways in which we may move beyond either-or dichotomies such as emotion versus cognition, or cognitivism versus non-cognitivism, which might hinder progress in our understanding of moral judgement.

2. Moral Judgement as Bayesian Inference

The psychological mechanism of moral judgement can be described by appealing to three neurocomputational ingredients:

- The prior representation of social norms concerning socially/morally (in)appropriate or right/wrong behaviour in a given context.
- The relationship describing how likely it is that any moral judgement gives rise to certain sensory data.
- The continuous updating of norms courtesy of Bayesian inference.

The first two ingredients are relevant to explaining especially moral judgement as a state of the mind. The other helps us pick out important features of moral judgement as a process.

With these ingredients in place, making a moral judgement would amount to activate what can be called ‘norm priors’ and combine the information carried by norm priors with incoming sensory data. I shall now explain the three ingredients.

2.1 Moral Judgement as Prior

Agents’ knowledge about how one ought to behave under certain types of circumstances is a subset of their social and moral knowledge. For example, an agent’s knowledge that she ought to buy the next round of drinks at the pub is a subset of her social and moral knowledge. As the previous Chapter suggested, social knowledge can have multiple formats. It can be encoded as clusters of prototypes, exemplars, scripts or rules in a system. In Chapter 2 it was argued, furthermore, that one fruitful way to conceive of what it is to have beliefs and preferences is in terms of probability distributions encoded by the nervous system. If agents' social and moral knowledge is constituted by their social/moral beliefs, then one fruitful way to conceive of social and moral knowledge is in terms of probability distributions. Now I articulate a few points made in Chapters 1 and 2 and relate them to moral judgement.

A probability distribution, recall, describes the range of possible values that a random variable can attain and the probability that that variable is within some range. I call *social distributions* those distributions that encompass social random variables. Social random variables describe features relevant to interact appropriately with others. Such features can be facial expressions, motion dynamics, eye gaze direction, and ostensive social signals such as certain types of gestures or tones of speech used (typically deliberatively) to communicate determinate intentions. Each social random variable—for instance, a facial expression—can take different values—for instance, a facial expression can be sad, angry or happy.

There can be correlations among social variables, or among specific values of social random variables. For instance, a certain facial expression may be correlated with particular motion dynamics; or a particular tone of voice may be correlated with a certain posture and certain ostensive social signals. By long association, we can expect many social features and events to be almost always together in certain types of circumstances. The social distributions describing such features and events, which our cognitive system might encode, can be joint over multiple variables. Depending on the details of one's learning trajectory, correlations among different social random variables and among different values of different social random variables have varying strength.

It is worth recalling that if our cognitive system encodes multivariate social probability distributions, then we need *not* have infinite representational resources or infinite information processing capacity, since probability distributions can be represented with small sets of values—for example, it suffices to represent a multivariate normal distribution with its mean and covariance matrix—and transformations of such distributions can be carried out by algorithms that approximate exact Bayesian computations. For example, Monte Carlo or stochastic sampling-based approximations of Bayesian computation are algorithmic schemes, which neural activity might implement in feasible ways (see e.g. Fiser et al. 2010).

Agents' social/moral knowledge—I suggest—is built on the multivariate social distributions encoded in a hierarchical way in their cognitive systems. From multivariate social distribution, agents would have knowledge about, for example, how one is expected to behave in a given situation, how other people are likely to react to one's behaviour in some type of circumstance, how one will react to certain behaviour displayed by others or by her.

Social/moral knowledge, on which our moral judgements depend, might consist of distributions over a range of candidate hypotheses, which specify how one ought to behave across different types of situations in the social environment. Social distributions encoded in an agent's cognitive systems specify the strength with which the agent entertains any hypothesis *before* any observation about the hypothesis is available. The subset of social distributions corresponding to such *a priori* mental states about (in)appropriate or right/wrong behaviour in the moral/social environment can be called *norm priors*.

The moral qualities attached to any norm prior depend on representations at still higher levels in the hierarchy of distributions in our cognitive system. Such representations constitute a bedrock of general knowledge concerning how our attitudes and our actions should take into account the needs, the desires and expectations of others. The *importance* (or value) attached to a behaviour or an action is updated courtesy of reward-prediction errors, which are mainly, but not exclusively, triggered by the observation of the sensory- and reward-data given rise by that behaviour. More on this in section 2.3 below.

The normativity of moral judgement depends on this bedrock of value-knowledge which infuses the world with value, with importance. Value-knowledge provides us with general moral convictions and moral concerns. It guides our behaviour by specifying goals we deem important, tracks changes in our motivational states and causally affects our normative judgements. This bedrock is shaped by the workings of value-based systems like the RL-systems described in Chapter 1. These systems not only enable smooth, adaptive interactions with others, but, as Chapters 5 and 6 will suggest, underlie our capacity to care about things and to create importance in our world as well. Let me now characterize norm priors more precisely.

Norm priors can be formalized thus:

[1] *Prob* (*A* ought to φ in *S*),

where *A* is some agent, φ specifies an action and *S* is a type of situation. In our cognitive hierarchy, distributions of form [1] would be encoded at higher levels than distributions constituting our bedrock of value-knowledge. These distributions have the following form:

[2] *Prob* (φ has value V | social representations R_s),

where R_s is the representation of situation S. At still lower levels we would find social distributions of the form:

[3] *Prob* (R_s | situation S).

Further down the hierarchy would lie encodings of objects, shapes, colors, textures, sounds, and so on, until we reach encodings of simple physical quantities such as velocity or orientation.

With this characterization in hand, I put forward the hypothesis that moral judgements as states of the mind can be fruitfully understood as norm priors encoded in our cognitive system. I now give some flesh to this hypothesis by highlighting four properties of moral judgements, which can be naturally accommodated if we understand moral judgements as norm priors.

2.1.1 Moral Judgement. Prior to What?

Norm priors are not prior to any experience or skill relevant to (in)appropriate or right/wrong behaviour. In general, priors refer to a learner's degree of belief in a hypothesis *before* observing data relevant to that hypothesis in the situation at hand. This does *not* mean that the learner's prior refers to her degree of belief in a hypothesis before she has acquired any body of knowledge or skill relevant to make a judgement or to act in the situation at hand. An agent's norm prior, for example, might refer to her degree of belief in the hypothesis that she ought to buy the next round of drinks in that situation *before* she observes the sensory data given rise by that behaviour in that situation.

The agent's norm prior reflects her state of knowledge and practical competence as she faces a certain situation. The agent judging that she ought to buy the next round of drinks may in fact have much relevant prior experience with social behaviour in pubs. She may have gained this body of knowledge from direct apprenticeship or from testimony, by reading books or listening to some friends' stories. This kind of prior experience is necessary in every aspect of ordinary moral affairs requiring some learned moral skills.

An agent's norm prior, therefore, does *not* consist in a hypothesis about what one ought to do which the agent entertains "at the beginning of the beginning," before the agent has undergone any experience or developed any skill (Suppes 2007). Rather, it is almost always the case that there has been some experience and that some skills have been developed prior to the elicitation of a norm prior in a particular situation.

Norm priors can accommodate that it is almost always the case that our moral judgements obtain against a background of moral experience and skills gained during continuous social apprenticeship. I shall have something to say about "the beginning of the beginning," about which types of norm priors might be hardwired in our cognitive systems in section 3.2.1 below.

2.1.2 Three Gradable Properties of Moral Judgement

In general, the more spread out a random variable of a probability distribution, the greater the entropy of that distribution, and the greater the uncertainty the agent has towards the corresponding hypothesis (Kruschke 2008b). Both the value-knowledge and the social distributions encoded in an agent's cognitive system have varying degrees of uncertainty (or entropy). So, norm priors have varying degrees of uncertainty (or entropy). This can naturally accommodate the fact that moral judgement seems to have a number of “gradable” properties (on these properties see Smith 2002).

The first property, which moral judgement seems to share with all judgement, is the level of confidence or uncertainty that an agent has that a behaviour in some situation is good (or bad), right (or wrong), appropriate (or inappropriate) as she judges it to be. Agents may be more confident that stealing is wrong than they are that not buying the next round of drinks at the pub is wrong.

The second property of moral judgements is that they are more or less stable in the face of new information. This feature seems to apply generally to all judgement as well. Agents, for example, may be equally confident that stealing is wrong and not buying the next round of drinks at the pub is wrong. But, in the face of incoming new information, agents’ confidence in the former judgement is more stable than their confidence in the latter judgement.

The third property, which seems specific to normative judgement, is the degree of importance, or value that an agent assigns to some behaviour in some situation. Agents, for example, can assign high value (or high importance) to not to steal, but they can assign higher value to not to murder.

As pointed out by Smith (2002, Sec. 2), these three properties are relevant to explaining action. The more confidence agents have that they ought to do something under some type of circumstance, the more they will be motivated to do it, all else being equal. The more value they assign to certain behaviour under some type of circumstance, the more they will be motivated to do it. Over time, the motivation that agents have to engage in some type of behaviour in some social situation co-varies with the level of stability of their judgement concerning that type of behaviour.

If we understand moral judgement as norm prior, then confidence in one's moral judgement corresponds to the entropy of the underlying norm prior. So, the higher the entropy of an agent's norm prior, the less certain the agent that she ought to do something, as specified by the norm prior.

Moreover, by considering that norm priors are constituted by a hierarchy of representational levels, with representations of value at higher levels, we can explain situations where agents are confident that some type of behaviour has certain social features, but they are less confident about the moral qualities of that type of behaviour.

Different distributions in the hierarchy can in fact have different levels of entropy. In particular, the entropy associated to some distributions underlying our body of value-knowledge can have high entropy, while social distributions at lower levels in the hierarchy might have lower entropy. As we learn how to successfully navigate our social space, and how to appropriately judge situations in our social/moral environment, the uncertainty (or entropy) of our norm priors decreases. As the uncertainty of an agent's norm prior becomes lower and lower, the agent will be more motivated to engage in the type of behaviour specified by the norm prior, all else being equal.

2.1.3 Bias and Moral Judgement

One of the most robust findings in moral psychology is that moral judgement and social behaviour, more generally, can be affected by morally irrelevant situational factors (for reviews see Bargh and Williams 2006; Sinnott-Armstrong 2008). Here are a couple of examples. People's tendency to cheat or to act selfishly increases if they wear sunglasses or they are placed in a dimly lit room (Zhong et al. 2010); people's moral judgement is less severe after they wash their hands with soap and water (Schnall et al. 2008).

Situational factors bias moral judgement in that they incline an agent to make one judgement, or a decision, over another. As such, 'bias' does *not* entail a deviation from a normative standard of judgement. Situational factors trigger specific informational processes, which lead agents to put more weight on certain sources of information, to prioritize some representations at the expense of others, and—relevant to our topic—to activate some social distribution over others.

Because of such biases, people's moral judgements vary across contexts even if their body of relevant moral knowledge and skills remain constant. Some biases influence the ease of retrieval of relevant information, or make available counterfactual alternatives to a given hypothesis, or make us focus only on particular features of some situation. Biases affect informational processes underlying distinct aspects of practical reasoning dependant on memory or attention (see Sunstein 2005).

If an agent's moral judgement is understood as norm prior, then which norm prior is active in a given situation will depend on the factors present in the situation. These factors will bias the transformations carried out by our system along the hierarchy of social distributions leading to the construction and activation of norm priors. When two priors fit the sensory data equally well, biases are the only basis for deciding between them.

2.1.4 Verbalization

Finally, it should be clear that *some* of the information carried by norm priors is available to consciousness and can be verbalized, expressed in utterances and sentences. The overwhelming bulk of our moral knowledge, however, might not be accurately or approximately verbalized, as people often have beliefs and attitudes they are not aware of having. Yet this knowledge can affect our moral behaviour. Here is an example: In spite of their self-reported beliefs and attitudes towards black people, European American and African American are more likely to misidentify a harmless object as a gun if they are first shown a picture of a black man rather than a picture of a white man (Payne 2006).

If moral judgement is understood as norm prior, and norm priors are the kinds of probabilistic internal representations that the previous chapters have described, then we can accommodate the fact that many of our moral judgements are not introspectable or conscious mental states. Within the neurocomputational framework I embrace, beliefs and desires are understood as probabilistic representations. They need not refer to mental states of which we are aware or which we can verbalize. Yet, probabilistic representations such as norm priors have effects on agents' behaviour and can be controlled and manipulated. From this perspective, people can have some moral judgements, which they consciously "disbelief." To put it another way, what people can have access to and verbalize is just a little, often inaccurate, and approximate portion of the rich body of moral and social knowledge encoded in their cognitive system.

By understanding moral judgement as norm prior, we can naturally separate questions about how moral judgements develop and affect our behaviour from questions about how much we can verbalize of our moral judgements. With this separation, we can consider whether and how the moral judgements that are verbalized—or moral discourse more generally—can be used in important ways or can have some impact on moral thought. The relationship between language and moral thought will be the topic of the next Chapter.

2.2 Likelihood of Moral Judgement

The relationship describing how sensory data vary with any moral judgement can be called 'likelihood function of moral judgement.' Call 'Sensory Input *I*' the sensory consequences given rise to by the moral judgement '*A* ought to φ in *S*,' the likelihood functions of moral judgement can be formalized thus:

[4] *Prob* (Sensory Input *I* | *A* ought to φ in *S*).

This quantity is a function of both observed sensory data and moral judgement. The likelihood of a moral judgement is the probability of sensory input given the hypothesized moral judgement. It measures how expected some set of sensory inputs is for different moral judgements: it expresses to what extent a moral judgement fits some set of sensory inputs. Likelihood functions of moral judgements can be regarded as generative models of observing sensory input *I* under the hypothesis that one ought to φ in *S*. The likelihood of moral judgement reflects how probable it is that we receive, for example, the current sensory input given the judgement we entertain that one ought to take vengeance, or given the judgement we entertain that one ought to forgive.

It is noteworthy that the notion of likelihood is distinct from the notion of probability. Mathematically they are directly related to each other: The *likelihood* of some particular moral judgement given some observed dataset of sensory inputs (I_1, I_2, …, I_n) is equal to the *probability* of the observed dataset given the moral judgement. For example, the likelihood of the judgement that one ought to buy the next round of drinks at the pub given such sensory inputs as a smile, a pat on the back, a 'cheers' is equal to the probability of those sensory inputs, given the moral judgement that one ought to buy the next round of drinks at the pub. But likelihoods and probabilities differ in what they represent. For probabilities, the hypotheses (or parameters) are known and the data are unobserved. For likelihoods, the data are observed and the values of the hypotheses (or parameter values) are unknown. So, for likelihoods of moral judgement we don't know which particular moral judgement obtains in a particular situation. Likelihoods specify the probability of sensory data we receive given different possible moral judgements we could entertain.

Likelihoods of moral judgement can be relevant to describe the psychological mechanism of both moral judgement as a mental state and moral judgement as a process. On the one hand, the likelihood of norm compliance specifies how observed data are related to different moral judgements, understood as states. On the other hand, according to the view I favour, moral judgement as a process consists in combining norm priors with likelihoods. The next subsection will articulate this latter point. Let me expand on the former now.

Likelihoods of moral judgement are sensory estimates, which are relevant to understand the psychological mechanism of such aspects of our social and moral life as trusting, hoping and promising. Many of our moral judgements depend on trust and hope since they depend on trusting in the testimony of others or on hoping that something will be the case. For example, my judgement that one ought not to leave a tip in restaurants in Japan depends solely on trust in the testimony of others. Many of our moral judgements, furthermore, concern trust, hope and promises. For example, we may judge that we ought to keep promises or ought not to betray those who trust us. It should be clear that thus trusting, promising and hoping allow us to form relationships with others. We may depend on these social relationships to satisfying our needs, our desires and to accomplishing the projects we consider important.

But trusting, hoping and promising involve uncertainty. We are uncertain, for example, that people we trust will not betray us. We are uncertain that this person is trustworthy, and therefore that we should trust her. If it were certain that some people would pull through without betraying us, then it would be unnecessary to trust them. If it were certain that something will be the case, then it would make no sense to hope for it. If there were guarantee that people keep their word, then we would have no need to make promises.

Likelihoods of moral judgement provide us with information about this uncertainty, as they specify the relationship between moral judgements and the sensory inputs they give rise to. Likelihoods of moral judgement describe how our moral/social environment changes so as to produce sensory inputs from different possible moral judgements—or from behaviour conforming to a certain judgement. Ray et al.'s (2009) study on the Trust Game, which was described in Chapter 2, clearly illustrates this point. In their study, agents' likelihood functions specify the probability of observing a sequence of sensory data (e.g. the opponent's observed actions), given the hypothesis that the opponent is of a certain type. That reflects—to repeat—how probable it is that we would observe the opponent's current action, given that the opponent is trustworthy, or given that the opponent is shady. If opponents are in fact of a certain type, then a player may hold the judgement that she ought (or ought not) to trust them since that player's trust would successfully target trustworthy agents.

Likelihood functions of moral judgement, then, might enable agents to judge who is trustworthy, and therefore should be trusted, and to act on this judgement. Identifying how our cognitive system might encode and transform likelihood functions of moral judgement might shed new light on the psychological mechanism of at least trusting, hoping and promising.

2.3 Norm Update

We make a moral judgement—I hypothesize—by combining norm priors with likelihoods of moral judgements. More precisely, a moral judgement would be obtained by multiplying each norm prior by the value of the likelihood of moral judgement. At any given time, the moral judgement that we entertain is the least uncertain moral judgement, which is the peak of the posterior distribution *Prob* (*A* ought to φ in *S* | Sensory Input *I*). More formally:

[5] $$\textit{Prob}\ (A \text{ ought to } \varphi \text{ in } S \mid \text{Sensory Input } I) \propto \textit{Prob}\ (\text{Sensory Input } I \mid A \text{ ought to } \varphi \text{ in } S)\ \textit{Prob}\ (A \text{ ought to } \varphi \text{ in } S)$$

From [5], two points should be clear. First, when we make a moral judgement, that is, when we entertain the judgement that one ought to do something in a certain situation, we incorporate prior moral knowledge to estimates of the sensory consequences of possible moral judgements. Moral judgements are computed based on available moral knowledge and on incoming sensory input. Such computations might consist of Bayesian inferences carried out in the neural hierarchy underlying moral cognition. For any layer in the hierarchy, each posterior becomes a new norm prior and can be further updated based on incoming sensory input.

Second, even if we may not be aware of it, our moral knowledge is continuously updated based on new sensory information. The updating might be carried out courtesy of prediction-errors propagated along the hierarchy. Under the impact of new information, how stable the knowledge at each layer is depends on the shape of the distributions encoded at the immediate neighbour layers. The higher the entropy of some social distribution at some level, the more likely it is that the knowledge at that level will undergo revision. Let me illustrate the first point.[5]

Let i stand for the current sensory input; x_1 stands for a random variable describing the possible values of the feature computed by neural populations at layer 1 in the cortical hierarchy; x_h stands for all knowledge encoded at higher layers, e.g. contextual information about the social situation and more abstract value-knowledge. Neural populations at layer 1 come to represent the most probable values of x_1 by computing the *a posteriori* distribution that maximizes $Prob\ (x_1 \mid i, x_h)$. Assuming for simplicity that $Prob\ (i \mid x_1, x_h)$ does not depend on the higher-level information carried by x_h, we can say that the transformations brought about at layer 1 consist in multiplying the likelihood of x_1, $Prob\ (i \mid x_1)$, by the prior $Prob\ (x_1 \mid x_h)$. The prior carries information about the degree of compatibility of every possible value of x_1

[5] This illustration relies on Lee and Mumford (2003, Sec. 2).

with the high level knowledge x_h. The likelihood carries information about the impact of incoming input given any value of x_1.

According to this view, each cortical layer is an expert for inferring certain features of our social/moral environment. Populations of neurons at each layer in the hierarchy are mainly interested in the computations carried out by their immediate neighbours. Inference carried out by neural activity at one layer is constrained by both bottom-up data coming from the feed-forward pathway and the top-down information feeding-back.

Let's assume that our moral cognition is underlain by neural populations ordered hierarchically in four layers. Each layer computes a set of features with the top layer computing a moral judgement, or a "belief-state." Call these features x_1, x_2, x_3, the judgement j, and the incoming sensory input i. Each feature computed in the hierarchy is provided with a value-tag from the high-level body of value-knowledge. If we judge that an action is wrong or some behaviour is inappropriate, then it is wrong or inappropriate because of certain of its features and of their value-tags. Whether the behaviour is right/wrong or (in)appropriate is determined by the distributed probabilistic computations taking place along the hierarchy. Features represented by social distributions can be morally significant in some case, but can make a different moral difference in another type of circumstance. Features have variable moral relevance depending on the computations of other features of the case we face and on their value-tags.

Here is an example. We make the simplifying assumption that if in the sequence (i, x_1, x_2, x_3, j) any variable is fixed, then the variables computed at the immediate neighbour layers are conditionally independent. The moral judgement entertained by the system at a time, with incoming input *i*, can then be described with the multivariate distribution:

$$[6] \qquad Prob\ (i, x_1, x_2, x_3, j) =$$

$$Prob\ (i \mid x_1)\ Prob\ (x_1 \mid x_2)\ P\ (x_2 \mid x_3)\ Prob\ (x_3 \mid j)\ Prob\ (j)$$

From [6], it follows that the moral judgement entertained is computed thus:

$$[7] \qquad Prob\ (x_1 \mid i,\ x_2, x_3, j) \propto Prob\ (i \mid x_1)\ Prob\ (x_1 \mid x_2)$$

$$Prob\ (x_2 \mid i,\ x_1, x_3, j) \propto Prob\ (x_1 \mid x_2)\ Prob\ (x_2 \mid x_3)$$

And so forth until:

$$Prob\ (j \mid i, x_1, x_2, x_3) \propto Prob\ (x_3 \mid j)\ Prob\ (j)$$

Social feature x_1 is computed through activity in neural populations in layer 1. The computation of x_1 is affected by the bottom-up feed-forward data *i* and the probabilistic prior *Prob* ($x_1 \mid x_2$) fed back from layer 2. The feed-forward input drives the generation of a moral judgement; the feedback from higher layers provides the priors to constrain inference at lower layers. So the moral judgements, understood as mental states, we entertain at a time would just be the result of the interaction between these feed-forward and feedback signals. The least uncertain of our moral judgements is the moral judgement having more impact on our cognitive system and on our behaviour at a given time.

The second point highlighted by [5] is that moral knowledge is constantly revised and updated. There are at least three ways agents' moral knowledge undergoes changes. First, moral knowledge undergoes changes through conscious reflection. Agents reflect on what one ought to do in a given situation by assessing and weighing their reasons for behaving in a certain way rather than another. Conscious reflection on what one ought to do can take place within a dialogue with other agents, where different moral judgements are put forth for consideration and defended either by argument or by other persuasive means. This may lead to a revision of the shape of the distributions encoding the moral knowledge available for conscious reflection and verbalization. I shall get back to this point in the next Chapter.

The second way agents' moral knowledge undergoes changes is through random fluctuations in its underlying distributions or in the value-tags attached to features represented by such distributions. Ongoing brain activity is found over a wide range of spatial and temporal scales. The functional significance of variations in spontaneous activity is not clear, but it is not unreasonable to believe that it might be also associated with variations in our body of social and moral knowledge (for a review on the functional significance of ongoing activity fluctuations see Sadaghiani et al. 2010).

Agents' moral knowledge is constantly affected by the sensory input they receive. This is the third way it can undergo changes. An agent's moral knowledge will not undergo changes under the impact of sensory input at a given time only if the agent's moral knowledge at that time predicts exactly her incoming sensory input. If the agent's moral knowledge fails to predict the incoming sensory input at a time, then a prediction-error is triggered, which will lead to a revision of the body of moral knowledge. Prediction-errors would be the part of the feed-forward signal that is not predicted by the prior knowledge encoded in higher layers. In this case, the more prediction-error, the more our social/moral knowledge will be revised. Prediction-errors can bring about a revision of the value-tag attached to some social distribution or of the social distribution itself.

Here is an illustration of how new information can affect our moral judgement courtesy of prediction-errors. Other people’s utterances provide us with new information, which can impact our evaluative judgement. In an fMRI experiment, Klucharev et al. (2009) asked participants to rate the attractiveness of some faces while their brains were scanned. After each judgement, participants were informed about peers’ average rating. A conflict with the peers’ opinion elicited a response in the nucleus accumbens and the rostral cingulated zone similar to a prediction-error signal. The magnitude of this signal seems to have impacted on people’s evaluative judgement since it predicted conformity with peer rating. Participants, in fact, judged again the same faces outside fMRI scan after thirty minutes. Those initially in disagreement with the group rating tended to change their judgement toward conformity. Prediction-errors may trigger long-term conforming adjustment of an individual’s judgment.

3. Neurocomputationalism at Work on Moral Judgement

Thus far I have provided a description within a RL-Bayesian neurocomputational framework of some central aspects of the psychological mechanism of moral judgement.

It is now time to put this neurocomputational proposal at work and see what it can bring to the table. The remainder of the chapter argues that the neurocomputational description of moral judgement put forward above can shed new light on puzzling findings about specific patterns of moral judgement. I focus on the pattern of moral judgement displayed by psychopaths and small children.

3.1 The Moral Judgement of Psychopaths

Psychopathy is a developmental disorder that involves pathological social behaviour. Psychopaths habitually violate important norms in their society. They are glib, impulsive, irresponsible, manipulative, egocentric, callous, lack empathy and have shallow emotions (Hare 2003). Psychopaths also make abnormal moral judgements. In particular, they have serious difficulty in drawing the so-called "moral/conventional distinction," which I now introduce.

Most people treat judgements such as "You ought not to steal" differently from judgements such as "You ought not to leave a tip in restaurants in Japan." The former, people would say, concerns a moral norm, whereas the latter a social norm (or convention). Most people would judge that violations of moral norms like hitting another person are more serious than violations of social norms like speaking without raising your hand. They would also deem the normative force of moral norms as less dependent upon authority figures and upon other people's expectations than the force of conventions. So, it seems that moral violations can be characterised by their consequences for the liberty, wellbeing and welfare of others; violations of social norms (or conventions) can be characterised as violations of behavioural uniformities structuring social interaction within a given social environment.

There is good empirical evidence that the capacity to distinguishing between moral norms and conventions/social norms is central to the normal development of our normative competence (Turiel 1983). A number of psychological experiments, using what is known as the *moral/conventional task*, have been run across nationalities, cultures and ages to test putatively defining characteristics of moral norms and conventions. The task consists in presenting subjects with violations of prototypical moral norms as well as with violations of other norms, and asking them a series of probe questions (e.g. “How wrong is the behaviour in the example?”; “Would the behaviour be wrong if some authority figure permitted it?”; “Would it be wrong also in a different place or at a different time in history?”; “Why is the behaviour wrong?”). Healthy subjects distinguish moral and conventional violations from the age of 39 months (Smetana 1993). Psychopaths have difficulty in drawing this distinction (Blair 1995).

There are two main puzzling findings about psychopaths’ judgement in the moral/conventional task. First, psychopaths tend to judge all transgressions as cases of moral transgressions (Blair et al. 1995a, 1995b). Psychopaths judge that, for example, it is *not* okay that a schoolboy walks out of the classroom without permission even if the teacher says it is permissible to do so. Second, unlike healthy controls, psychopaths ignore considerations about victims’ welfare or social disorder when they justify why some action is wrong. Psychopaths and healthy controls—it is noteworthy—do not differ in the way they draw, and justify, the moral/conventional distinction when they are confronted with positive acts like comforting a friend or wearing the uniform at school (Blair et al. 1995b).

3.1.1 Not by Emotion Alone

Authors like Jesse Prinz (2007) and Shaun Nichols (2004) link psychopaths’ incapacity to make judgements concerning moral and conventional/social norms to their emotional abnormality. Many discussions of psychopathy, in fact, identify lack of sympathy and incapacity to feel guilty as the deficits at the root of this disorder. Psychopaths appear to be indifferent to the concerns and some feelings of the others.

For example, in response to cues of distress such as the facial expression of a crying child they show no affective response (Blair et al. 1997). Psychopaths feel little remorse when they break moral or social norms. They show an incapacity to attribute guilt to violators of moral norms (Blair et al. 1995a). “The moral blindness of psychopaths—Prinz (2007, p. 46) writes—issues from an emotional blindness.” But the evidence does not warrant such a conclusion.

Psychopaths’ emotional profile is not flat: children with psychopathic tendencies, and adult psychopaths alike, are normal in their attributions of happiness, embarrassment and sadness to people described in short vignettes (Blair et al.1995a). So they don’t seem to have a general inability to experience emotion (Blair 1997). Psychopaths “show reduced skin conductance to sad, but not angry expressions. Moreover, children with psychopathic tendencies have been found to show selective recognition difficulties for sad and fearful expressions but not for angry, disgusted, surprised, or happy expressions” (Blair et al. 2001, p. 493). Psychopaths, thus, seem to be impaired in specific forms of emotional processing: they are probably impaired in emotional learning based on fear conditioning, and, as noted, in attribution of guilt. Although they show reduced emotional response in anticipation of punishment, they have normal response to reward (Blair et al. 2005). So, an appeal to an impairment in emotional processing might not suffice to explain psychopaths’ idiosyncratic pattern of normative judgement.

Neuropsychological research indicates that dysfunction in the amygdala is reliably associated to psychopathological behaviour (Blair 2003). The amygdala is one brain region most implicated in antisocial, aggressive and psychopathic behaviour (Raine and Yang 2006). Psychopathic individuals show a pattern of functional impairments generally displayed by patients with a lesion or dysfunction in the amygdala: They have "deficits in aversive conditioning, the augmentation of startle response by visual threat primes and fearful expression recognition" (Blair 2007, p. 388). More relevant here, psychopaths also show reduced activity in the amygdala during moral judgement (Glenn et al. 2009).

So, although "psychopathy is not associated with a lesion to a particular region, nor have all functions mediated by any particular region been shown to be compromised" (Blair 2007, p. 388), if psychopathy is reliably associated to a dysfunction in the amygdala, then one way to make progress in understanding psychopaths' abnormal pattern of moral judgement is by identifying possible computational roles of this brain region.

3.1.2 The Amygdala as Uncertainty-Detector and Psychopathy

Here is my proposal. The general computational roles of amygdala activation might be twofold. On the one hand, the amygdala would contribute to detecting the uncertainty associated to the structure of a situation with respect to a probabilistic model of that situation. That is, amygdala would detect unpredictable or uncertain situations. On the other hand, given its detection of the uncertainty of the situation, it would signal a need to learn: Having computed that a situation is uncertain to a certain degree (or unpredictable), the amygdala would bias an organism towards greater sensitivity to the causal and reward structure of the environment.

In the case of moral judgement, amygdala activation would contribute in the detection of the level of uncertainty underlying a given moral situation in terms of the variance of the posterior of moral judgement activated by that situation. The more uncertain the situation, the more active the amygdala will be. One way to capture the hypothesis that the amygdala might be sensitive to morally uncertain situations is in terms of a norm prior with maximum entropy, so that for any agent different actions seem equally (in)appropriate (or right/wrong) in that situation. Another possible way is in terms of a (nearly) "flat" (or constant) likelihood of moral judgement, such that different moral judgements fit equally well the sensory data, and so the cognitive system lacks information for selecting between competitive moral judgements.

The uncertainty underlying a moral situation might in turn be due to uncertainty with respect to the reward or the causal structure of a given environment. Having detected that a situation is morally uncertain, the amygdala would signal a need to learn about its structure, so that the entropy of the active prior could decrease and moral uncertainty resolved. I now make clearer and give some support to this hypothesis by describing Herry et al.'s (2007) experiment, which expands on findings about the amygdala in associative learning (for a review of amygdala functions see LeDoux 2008).

Herry and colleagues (2007) used a translational approach in humans and mice to ask whether the amygdala is essential "for processing sensory information that does not allow an exact prediction in time" (p. 5958). In their study, humans and mice were exposed to sound pulses. There was nothing specifically social or emotional about the pulses which were not associated with any other stimuli either. Herry and colleagues used two sound pulse sequences. One sequence was randomized so that the pulses occurred unpredictably at a variable interval. In the other sequence the sound pulses occurred predictably (every 200ms). It was found that a sequence of unpredictably-timed sound pulses was associated to sustained amygdala activity both in mice and humans, measured as c-fos changes in the mice—where c-fos is a protein used as an indirect marker of neural activity—and fMRI responses in the humans.

This finding supports the hypothesis that uncertainty *per se*, rather than emotional or social dimensions of stimuli, is sufficient to engage amygdala processing. Thus, amygdala activity might be tuned to the level of uncertainty of the statistical structure of the environment. Recall that this level of uncertainty depends on the probabilistic model of the environment encoded in agents' cognitive system. Amygdala activity, then, might contribute to the detection of whether there is a significant mismatch between the model and the structure of the environment an agent finds herself. More specifically, it might "track a quantity, known as associability, which reflects the extent to which each cue has previously been accompanied by surprise (positive or negative prediction errors" (Li et al. 2011, p. 1250).

One possible function of such uncertainty-detection could then be connected to learning about the environment. Amygdala-based computation of uncertainty might bias the agent towards greater vigilance to the contingencies in that environment (Blackford et al. 2010). More specifically, amygdala activity might control "learning rates dynamically, accelerating learning to cues whose predictions are poor and decelerating it when predictions become reliable" (Li et al. 2011, p. 1250).

The second part of Herry and colleagues' study addressed the hypothesis that amygdala-based computations supports important aspects of learning about environmental contingencies. They asked what behavioural effects amygdala response to uncertainty could have. In this second part, both humans and mice were engaged in tasks indexing stress and anxiety while one of the two sound pulse sequences played in the background. In particular, the human subjects were engaged in a dot-probe task where they viewed angry and neutral faces on a screen and had to press a button when a dot appeared in the location previously occupied by the face.

Compared with the predictable tone condition, in the unpredictable tone condition both mice and humans behaved more like anxious, hyper-vigilant subjects. Specifically, compared with the predictable tone condition, human subjects showed shorter reaction time when the dot occupied the position of the angry face instead of the neutral face when exposed to a sequence of unpredictable tones. Thus amygdala activity in response to an uncertain sensory environment seemed to bias responses towards greater sensitivity to threats, or more generally to biologically-relevant stimuli. By enhancing vigilance, amygdala activity might then signal the need to learn the structure of the environment (Whalen 1998). The role of the amygdala as uncertainty-detector might be more fundamental than—or even account for—its involvement in emotion and social cognition (Pessoa and Adolphs 2010).

3.1.3 What's Wrong with Psychopaths' Moral Cognition?

If psychopathy is a developmental condition, then psychopaths' abnormality is probably linked with moral judgement as a process. Psychopaths would have moral knowledge but they would be incapable of updating it. The moral cognition of psychopaths might be deviant, first and foremost, as a result of an insensitivity to the uncertainty of a given situation measured with respect to an internal probabilistic model of the situation. Because psychopaths would be unable to detect morally uncertain situations, they would lack signals enhancing their vigilance and disposition to learn about the structure of a given social situation.

Given the wealth of evidence indicating that amygdala activity can facilitate memory consolidation in other neural structures, sustained amygdala activity associated with detection of morally uncertain situations "may represent one possible mechanism by which prediction errors generated in one brain area may influence more widespread memory systems" (Herry et al. 2007, p. 5965). So, psychopaths' capacity to update and revise their moral knowledge might be compromised due to a lack in some types of prediction-errors or because the prediction-errors they generate fail to influence storage of new moral information.

Why Do Psychopaths Treat Conventional Wrongs As If They Were Moral Wrongs?

Psychopaths treat conventional wrongs as if they were moral wrongs because they have difficulties in updating their moral knowledge, in particular their value-knowledge. Such difficulty would ultimately depend on their blindness to uncertainty underlying social situations, which makes it difficult for their internal models of the social environment to be updated.

Abstract norms prohibiting harmful and unjust behaviour might be the norm priors hardwired in our cognitive system which constitute our moral knowledge at "the beginning of the beginning"—more on this in a moment. We would revise this body of "prior" moral knowledge and pick up other types of norms such as the local social norms and conventions structuring our social environment by learning from repetitive social interaction or by being explicitly instructed.

Because psychopaths are blind to uncertainty, they will have difficulties in learning from direct interaction with their social environment. Given that detection of uncertainty typically triggers heightened vigilance towards biologically-relevant stimuli, towards threats in particular, psychopaths may be less vigilant to social punishments. Thus, psychopaths would not be able to update their moral knowledge via value-based learning, in particular via social-punishments. They would not be able to revise their normative knowledge on the basis of the reward-consequences of their behaviour either.

Explicit instruction can enable them to pick up at least some of the social norms and conventions that regulate interaction in their social environment. This type of learning may be insufficient, however, to convey the gradable properties of moral judgement since, in general, explicit instruction conveys information about rules as if they were exceptionless, non-gradable generalizations (Cf. Rogers and McClelland 2004). But conventions, unlike moral norms, are usually not treated as exceptionless generalizations. So, although along their learning trajectory psychopaths can acquire moral norms and conventions, the way they learn about them does not allow for distinguishing between the two types of norms. Hence they make judgements about conventional violations as if they were moral.

Why Are Psychopaths Blind to the Welfare of Victims of Norm Violations?

If psychopaths have a difficulty in learning about the structure of their social environment, then they will tend to have difficulties in representing reliably how sensory input varies with any moral judgement. This is partly because of psychopaths' lack of vigilance towards the sensory consequences of any norm-compliant behaviour. If they are not vigilant in this sense, then the sensory consequences of any norm-compliant behaviour will provide unreliable data about any of their particular moral judgement. In other words, psychopaths' likelihood function of moral judgement is "wide," and consequently their moral judgement will be more strongly influenced by their norm prior. Being more strongly influenced by their norm priors, psychopaths will tend to put less weight on the consequences of a given moral or conventional transgression when they are asked to justify their judgement. Hence, when they justify their moral judgements, psychopaths will make predominant reference to information encoded in the norm prior, that is, to information about the norm itself (e.g. "It is not acceptable to do that"); compared to healthy controls, they will be less likely to make reference to other's welfare ("Doing that will hurt that person") or to the disruption caused by the transgression ("The class will be distracted if I do it").

Four Predictions

Specific predictions can be drawn from this diagnosis. First, if psychopaths are insensitive to the uncertainty of a moral situation, then, in comparison to non-psychopathic subjects, they will show less cognitive dissonance and less anxiety in judging a potentially problematic moral scenario, as they will be less prone to moral uncertainties. If amygdala activation detects uncertainty, and psychopaths are insensitive to the uncertainty of a moral situation, then given a moral scenario, psychopaths' norm priors will generally display a smaller degree of entropy than non-psychopaths' norm priors, or their likelihoods of moral judgements will generally be more "peaked" than non-psychopaths' likelihoods of moral judgements.

Second, psychopaths will be less vigilant—that is, they will show higher sensory thresholds throughout sensory cortex—when they have to judge a moral scenario; in particular, they will be less vigilant towards potentially-threatening stimuli in uncertain moral situations.

Third, psychopaths will be less disposed to revise their body of moral knowledge in comparison to non-psychopaths in the face of new relevant moral information. This learning impairment might be due to lack of prediction-errors or to a failure to consolidate memories courtesy of prediction-errors.

Fourth, if explicitly instructed about the authority-dependent nature of specific conventions, psychopaths will tend to show a normal capacity to draw the moral conventional distinction.

Rather than depending on an emotional deficit, therefore, the deficiencies in moral judgement of the psychopath might perhaps be consequence of a learning deficit, which ultimately would be consequence of an incapacity to detect and deal with uncertainty in morally significant situations.

3.2 Children's Moral Judgement

Children as small as three years of age can distinguish between moral norms (e.g. norms involving justice and harm) and conventions. So, small children seem to be equipped with abstract information about the moral world. At the same time, children learn about their social and moral environment from direct experience.

The type of information they acquire during development might not be sufficient to ground the moral/conventional distinction. The explicit moral education that children receive seems to be mainly directed towards social norms rather than moral ones. Casual observation suggests that most advice and corrections that children receive from caregivers and parents are directed towards social norms or conventions (“Don’t burp!”, “Wait for your turn!”) rather than moral norms (“Be just!” or “Don’t kill your mates”). Where does small children’s moral knowledge come from?

3.2.1 Children as Probabilistic Learners and Their Built-In Priors

I wish to suggest with Gopnik et al. (2010, p. 342) that “the child is a probabilistic learner, weighing the evidence to strengthen or reduce support for one hypothesis over another.” From this perspective, we might explain findings on small children’s judgement in the moral/conventional task by appealing to “evolutionarily built-in” norm priors.

One argument in support of this hypothesis goes like this:

P1. At least partly, natural selection has shaped our psychological tendencies.

P2. Humans are generally averse to risky states—where objective probabilities are known—and to uncertain states—where objective probabilities are missing.

P3. Risk aversion is evolutionary advantageous under many circumstances.

P4. If risk aversion is evolutionary advantageous under many circumstances, then, *a fortiori*, uncertainty aversion is evolutionary advantageous under many circumstances.

P5. Actions that involve harm and injustice bring about highly uncertain social states.

C. Therefore, humans may have inbuilt priors such that they averse to actions that involve harm and injustice.

I take it that P1 is non-contentious. P2 is underwritten by a substantial wealth of evidence (see e.g. Weber and Johnson 2009). P3 and P4 can be justified by appealing to Friston's free-energy principle, which we encountered in Chapter 1, thus. In an evolutionary setting, "model"-selection (or agent-selection) is constrained by free-energy minimization: models with the lowest average uncertainty are the ones who are likely to survive and passed on to the next generation.

P3 can receive independent justification as well. Here I draw on Samir Okasha (2007). Okasha's argument is that, under realistic assumptions, types of organisms with a lower variance in their reproductive output—individuals who are risk-averse with respect to their offspring—have fitness advantage under a variety of circumstances. Suppose—Okasha argues—that there are only two types of organisms in a population. They reproduce asexually, and their types are transmitted faithfully from parent to offspring. Type A organisms have fixed reproductive output (e.g. 5 offspring); type B organisms have a reproductive output that varies stochastically (e.g. 10 or 0 offspring with 0.5 probability). Although both A and B organisms have the same expected number of offspring E (E(B) = 0.5 * 10 + 0.5 + 0 = E(A) = 5), they do not have the same expected frequency of offspring. The frequency of a type X after a generation is given by:

NUMBER of Offspring of Type X / TOTAL number of Offspring in the Population

In a population with 2 organisms, one of type A the other of type B, after one generation, the population will contain 5 A and 10 B with 0.5 probability and 5 A and 0 B with 0.5 probability. By applying the formula above, in the former case the frequency of the A type will be 1/3, in the latter 1. If we compute the expected frequencies F, we have:

$F(A) = 0.5 * 1/3 + 0.5 * 1 = 2/3$; and

$F(B) = 0.5 * 2/3 + 0.5 * 0 = 1/3$.

Type A has higher expected frequency in the second generation. Since frequency is what matters for evolution, type A is fitter than B because its lower variance in reproductive output, or, put it in other words, because it is averse to risk. Therefore, under many circumstances evolution seems to favor risk-averse organisms.

If P3 is plausible, then P4 is plausible *a fortiori*. The consideration in support of P4 is simple. What is generally referred to as a 'risk' involves knowledge of objective probabilities. In the case of 'uncertainty,' instead, objective probabilities are unknown: they have to be guessed on the basis of prior experience. If risk-averse organisms have an evolutionary advantage over risk-seeking ones in many circumstances, then uncertainty-averse organisms will have an evolutionary advantage over risk-averse ones, as dealing with risk is less computationally-consuming than having to deal with uncertainty. Uncertainty-averse behaviour may be evolutionary advantageous. And our tendency to be averse to uncertainty might have an evolutionary explanation.

P5 seems plausible as well. Types of behaviours that are harmless or that promote justice seem to be particularly efficacious to minimize agents' uncertainty. Social and moral institutions "create order out of chaos, […] make our lives more predictable, and thereby allow us to devote less of our resources to solving recurrent social problems repeatedly" (Schotter 1981, p. 143). Under most circumstances, behaviour that involves harm or injustice tends to bring about uncertain states. Breaking a promise or punching others are such types of behaviours: situations where injustice and violence are systematically pursued are highly chaotic. Compared to behaviour like keeping a promise, the sensory consequences of harmful or unjust behaviours seem to be relatively harder to estimate.

If my argument in this section is sound, then there are grounds to conclude that at "the beginning of the beginning" humans might have norm priors such that they judge behaviour involving harm and injustice as wrong. Such behaviours would be "wrong" partly because they are catalysts of uncertainty. We may have an evolved bias to avoid types of actions like breaking a promise or hurting others. This built-in norm prior might explain why small children can distinguish moral from conventional violations.

Conclusion

This chapter has expanded on the Bayesian Brain Hypothesis and brought the Bayesian framework to bear on moral judgement. Specifically, it has put forward the suggestion that some central aspect of the psychological mechanism of moral judgement can be described within a RL-Bayesian neurocomputational framework. It has argued that this framework promises to shed new light on puzzling findings about psychopaths' and children's patterns of moral judgement, thereby helping us to explain them.

CHAPTER 5.

Leges Sine Moribus Vanae.[6]

Does Language Make Moral Thinking Possible?

Does language make moral thought possible? After having put forth an account of moral judgement in the previous chapter, I now tackle this specific question. More generally, in this chapter I explore the relationship between language and moral cognition by engaging with some relevant aspects of Andy Clark's work. I also point to one important capacity, the capacity for *florid control*, which might be enabled by the workings of RL algorithms implemented by dopaminergic circuits.

Clark's unabashedly transdisciplinary work and argumentative style represent an ideal platform for advancing a debate such that concerning the relationship between language and moral cognition. By bringing insights and results from various disciplines to bear on the understanding of such a relationship, Clark claims that human language explains the possibility of moral thought. He argues for a supra-communicative view of language according to which we use language not only, or mainly, to communicate (Clark 1998; 2006a; 2006b). Language, for Clark, augments our cognitive abilities, makes learning easier, facilitates us to offload our memory, helps us to structure the environment where we live, to manipulate and re-organize complex data-sets. We also use language to coordinate our interactions, make plans, persuade others and simplify complex tasks. Courtesy of language we can access our own cognitive practices from a second-order stance. Language, importantly, appears to make possible new domains of thinking (Bermúdez 2003; Carruthers 2002; Clark 1997, Ch.10; Dennett 1991, Ch. 8).

[6] From Horace, *Odes*, III, 24. Transl.: "Laws without morals [are] useless".

For Clark, the presence of language and the way people use it mark a fundamental divide between humans and all the other animals with respect to *moral* cognition. In an exchange with Paul Churchland, Clark writes:

> Recent work in cognitive science highlights the importance of exemplar-based know-how in supporting human expertise. Influenced by this model, certain accounts of moral knowledge now stress exemplar-based, non-sentential know-how at the expense of rule-and-principle based accounts. I shall argue, however, moral thought and reason cannot be understood by reference to either of these roles alone. Moral cognition—like other forms of 'advanced' cognition—depends crucially on the subtle interplay and interaction of multiple factors and forces and *especially* (or so I argue) on the use of linguistic tools and formulations and more biologically basic forms of thought and reason (Clark 2000a, p. 267).

More recently, in a debate with John Haugeland, Clark (2002a) presses the same point by arguing that linguistic objects like labels, words and sentences radically transform and expand the cognitive space our minds can explore. In particular, language would make available to our minds a "social-normative space."[7] According to Clark, the unique profile of our moral cognition, which John Haugeland calls 'norm-hungriness', is a "secondary effect of getting language going" (Clark 2002a, p. 54, discussion).

I take issue with Clark and argue for two claims. First, language is probably not necessary for moral cognition: at bottom, moral cognition is probably a kind of

[7] Here I use 'moral cognition' and 'moral thinking' interchangeably, I also use 'social-normative space' as akin to 'moral space'. Although I acknowledge that there is a spectrum of social behaviours some of which tend to be called 'moral', for my purposes it is not necessary to precisely define 'morality'. My use of the terms furthermore is consistent with Clark's, Churchland's and Haugeland's whose works are my main focus.

skill dependent on basic capacities of pattern-recognition and social learning. Second, our unique *norm-hungriness* could depend on our capacity for *florid control* rather than on language. This capacity will be further explained in the next chapter, in relation to the capacity to care.

The chapter is in four sections. The first section gains a broader perspective on the topic by rehearsing some of the central ideas in Clark's exchanges with Churchland and Haugeland. Then, it reconstructs Clark's main argument for why the moral domain of thinking is made possible by the presence and use of language. The second section is in three parts. It firstly challenges Clark's view on the relationship between language and moral cognition. Secondly, it puts forward the hypothesis that norm-hungriness could depend on humans' capacity for florid control. Finally, it considers to what extent my disagreement with Clark is merely terminological. The third section provides a succinct map of different effects that language can have on moral cognition. The last section summarizes the claims made and defended and points to questions for further research.

1. Churchland, Haugeland and Clark's "Discursive Construction of the Moral Space"

I begin by introducing Paul Churchland's argument for why language is probably unnecessary for moral cognition (Churchland 1995, Ch. 6 and Ch. 10; 1996; 2000). As we have seen in Chapter 3, Churchland claims that moral cognition is a kind of perceptual skill based on pattern recognition and prototype-based learning and categorization. As such, its possibility would not depend on certain uses of language or on linguistically codified rules.

Churchland's argument—as already noticed—is deeply influenced by results from cognitive neuroscience and by the requirements of neural networks modelling. One of its premises is that artificial neural networks are especially relevant for understanding human and animal cognition.

If neural networks don't process information by relying on any system of linguistic symbols or linguaform rules, then cognition does not probably require any system of linguistic symbols or linguaform rules. Linguistic symbols or linguaform rules are not causally involved in neural networks' processes. Therefore, cognition does not probably require any system of linguistic symbols or linguaform rules, to the extent that neural networks are relevant to understanding it. Churchland concludes that "a normal human capacity for moral perception, cognition, deliberation, and action, has rather less to do with rules, whether internal or external, than is commonly supposed" (Churchland 1996, p. 101). The possibility of moral cognition would not depend on language.

Moral thought would rather depend on perceptual skills acquired over a life-time of social experience. For Churchland, we learn how to recognize a wide variety of situations and how to respond to them by relying on a library of moral prototypes that we acquire from interaction with others. Moral prototypes, recall, are statistical central tendencies extrapolated from concrete moral examples encountered in a variety of social interactions. They facilitate us to classify and comprehend new social situations, and to respond to them appropriately. The successful navigation of our social and moral environment would ultimately depend on our prototype-based perceptual or recognitional skills which are embodied in configurations of synaptic weights of appropriately trained neural networks.

Clark agrees with Churchland on the relevance of artificial neural networks and cognitive neuroscience for understanding cognition. He argues, however, that language plays a fundamental role in constituting some of our social behaviours as *genuinely* moral. For Clark, the presence of language and certain uses of language set us apart from other animals by making us genuine moral agents.

In reaction to Churchland, Clark emphasises two points. First, moral judgement, moral deliberation and social decision-making are capacities that have a fundamental communal and collaborative dimension. Clark notices that "missing so far from the discussion [on the foundations of moral thought] is any proper appreciation of the special role of language and summary moral maxims within a cooperative moral community" (1996, pp. 120-121).

Moral judgement has a fundamental cooperative dimension since it involves being sensitive to the needs, reasons and desires of others. Such sensitivity demands a "commitment to finding routes through the moral space that accommodate multiple perspective and points of view" (Clark 2000b, pp. 309-10). In order to find such "routes through the moral space" language is essential because it makes possible to us to give and share reasons for our behaviour (Clark 1996; 2000b).

Second, for Clark, the domain of moral thinking is made available to us by linguistic objects such as moral labels, codified rules and social classifications. Churchland maintains that the role of this linguistic, external scaffolding is to offload, preserve and share our moral knowledge; but language, according to Churchland, is not causally necessary to make the moral space available to us. For Clark, instead, without a linguistic apparatus we would be blind to those behavioural patterns and concepts in our life that constitute the moral domain of thinking. Thus, language, for Clark, is causally necessary (or constitutive) to moral thinking.

The conjunction of these two claims can be called the thesis of the *discursive construction of the moral space*. Clark's thesis can be summarized thus:

> the moral realm comes into view, and moral cognition is partially constituted, only by the joint action of neural resources we share with other animals and the distinctively human infrastructure of linguaform moral debate and reason (Clark 2000b, p. 311).

More recently, Clark has articulated this thesis during an exchange with John Haugeland (Clapin 2002, Part I). Haugeland holds that humans have a peculiar norm-sensitivity or insatiable norm-hungriness which is unique to our species and is prior to the development of our impressive linguistic abilities. Language, for Haugeland, could not get off the ground without this norm-sensitivity or norm-hungriness.

Although it's not clear what Haugeland intends exactly by 'norm-hungriness,' one plausible way to understand this notion is in terms of a need or desire to create and abide by a multitude of norms. Such a need would lead to societies where the "structures of a community can rely on the fact that almost all its members will abide by almost all the norms almost all of the time" (Haugeland 2002, p. 32).

According to Haugeland, our norm-hungriness would depend on some neural innovation: "The native wetware endowment of homo sapiens—Haugeland writes (2002, p. 31)—has to have evolved so as to support our norm-susceptibility and norm-hungriness." Such a neural innovation is distinct from any neural circuits implementing our linguistic capacities. It is the very possibility of language and language-use that depends on the neural circuits implementing norm-hungriness.

Haugeland suggests that norm-hungriness and social norms exert "a kind of 'normative gravity' [such that] … when an individual's dispositions stray from producing [normal, conformist] behavior … they are 'pulled back in'" (Haugeland 2002, p. 32). Such normative gravity would promote "tight clumps [of sociality] separated by large empty gaps" (Ibid.). It would promote, that is, the emergence of discrete, identifiable types of social behaviour. This, in turn, would pave the road to digitalness which seems to be a prerequisite for the emergence of language. For Haugeland, therefore, "social norms may have laid the groundwork for language in a more basic way … by enabling the digitalization of behavioral types" (Ibid., p. 33).

According to Haugeland's account, Clark would then be "norm-blind." Clark is focused on language; and language is the wrong target for understanding the rise of moral cognition since moral cognition and norm-hungriness would be prerequisite rather than consequences of human language. In fact, Clark's thesis of the discursive construction of moral cognition "depicts norm-sensitivity and norm-hunger as *secondary* effects of our linguistically enhanced capacity to target biologically basic processing resources on increasingly abstract and higher order domains" (Clark 2002a, p. 40).

Clark, in reaction to Haugeland, argues that it was the emergence of language that allowed us to objectify "complex features and relations" which "ma[de] available new, quasi-perceptual, spaces for reasoning" (Ibid., p. 42); and it was this phenomenon that cranked up the unfolding of the social-normative space that our minds can explore.

Let me now explain more carefully Clark's argument in support of his thesis of the discursive construction of moral cognition.

1.1 Chimps, Representational Re-Coding and Morals

In his exchanges with Churchland and Haugeland (Clark 1996; 2000a; 2000b; 2002a) Clark's main argument in support of his thesis of the discursive construction of the moral space is a study of analogical reasoning in chimps (*pan troglodytes*) by Thompson, Oden and Boysen (1997). Thompson and colleagues seemingly show that chimps trained to use arbitrary plastic tokens of different shapes and colors, which are consistently associated with pairs of identical objects (e.g. two shoes or two cups) or with pairs of different objects (e.g. one shoe and one cup), learn to grasp abstract relationships.

The task in Thompson and colleagues' study was to identify higher-order relationships of sameness or difference by exploiting the plastic tokens as stand-ins for same-relationship and difference-relationship. The chimps had to recognize a display of Cup/Cup as an instance of the same-relationship and a display of Cup/Shoe as an instance of the difference-relationship. Experience with the external plastic tokens seemed to be necessary for the chimps to solve more abstract problems. Presented with a display of Cup/Cup and Shoe/Shoe, or a display of Cup/Shoe and Cup/Shoe, the chimps could identify the higher-order relationship of sameness also in this case. Chimps were also able to recognize that a display of Cup/Shoe and Cup/Cup instantiated the higher-order relationship of difference. Chimps that did not undergo that kind of symbolic training could not succeed in the task. Hence it seems that prior experience with the external tokens was necessary before the chimps could perform successfully.

The conclusion drawn by Clark from this example concerns the capacity acquired by the chimps courtesy of *objectification* of abstract relationships—a process "akin to acquiring a new perceptual modality" (Clark 1998, p. 175). According to Clark, the chimps in Thompson and colleagues' study learn to solve a complex problem because of *representational re-coding*—which was mentioned in Chapter 2. Chimps would be able to symbolically re-code complex abstract relationships into iconically equivalent, simple, usable objects. Clark suggests that this capacity leveraged on the experience with the plastic tokens which enabled the chimps to acquire new mental representations. Such mental representations could stand in for the abstract regularities instantiated by the plastic tokens. Thus, for example, when the chimps faced a pair of identical objects, they could retrieve a mental representation associated to the same-relationship. When the chimps were confronted with two pairs of objects like Shoe/Shoe and Cup/Cup, they could retrieve and use two representations of plastic tokens of the same type. The task was thereby reduced to the first-order problem of recognizing that two tokens were of the same type. Let's now reconstruct Clark's argument with his case-study in hand.

Clark's first step consists in showing that words, labels and tags are tools that enable representational re-coding. Language, that is, enables us to objectify our thoughts and ideas in the same way experience with plastic tokens enabled the chimps to objectify abstract relationships. Language, according to Clark, would compress abstract regularities into basic cognitive objects. Such cognitive objects can make possible new domains of thinking. And in these new cognitive domains, the computational space we have to search in order to solve a certain problem would be dramatically reduced (Clark and Thornton 1997). These cognitive objects can enter a process of objectification themselves, thereby allowing us to create and explore further cognitive domains.

The second step in Clark's argument is important. Clark argues that moral cognition is a new domain of thinking made possible by the kind of representational re-coding enabled by language. Our use of public language enables us to compress abstract relations and features in cognitive objects anchored to moral labels and maxims. Normative talk would render certain features and abstract relationships visible and usable for us, just as experience with plastic tokens rendered abstract relations-between-relations visible and usable for chimps. The normative space of morality, according to Clark, is a virtual, higher-order cognitive realm; and norms, duties, rights, promises, commandments and obligations are examples of objects that populate such a realm. According to Clark, therefore, it is a process of objectification empowered by our use of language that makes moral thinking possible and builds up our unique norm-hungriness: our capacity to create, learn and act upon social norms.

I am not persuaded by Clark's argument. I agree with Clark that language plays an important role in the *unfolding* of the social normative spaces we inhabit, but I am not convinced by the second step in his argument. Specifically, I am not convinced that the kind of representational re-coding described by Clark as empowered by language is necessary for the rise of moral thinking and in particular for making possible a social-normative space. Human norm-hungriness does not probably follow from "getting language going." Our peculiar norm-hungriness would rather depend on *florid control*.

I am sympathetic, instead, with aspects of both Churchland's and Haugeland's accounts. On the one hand, I agree with Churchland that the very possibility of moral cognition does not depend on a linguistic apparatus. The account of norm compliance laid out in Chapter 1, in fact, does not make reference to language or linguistic resources. On the other, I am attracted by Haugeland's idea that humans' peculiar norm-hungriness might depend on some neural innovation.

2. Why Language Could Not Be Necessary for Moral Thinking

Is the capacity enabled by re-coding, and displayed by the chimps in Thompson and colleagues' study, causally necessary for creating and successfully navigating a social-normative space? To answer this question let's consider the case of macaque monkeys.

Macaque monkeys can judge whether two objects are identical on the basis of their physical features or of category similarity. They fail, however, in the type of high-order reasoning task where chimps can succeed (Thompson and Oden 1998). So, they seem to lack a capacity for representational re-coding as rich as that of chimps. How does this affect their capacity to navigate their social space? Is representational re-coding causally necessary for the rise of social norms and for making possible complex social interactions?

One famous study by Dasser (1988) with long-tailed macaques shows that the ability to recognize others' social relations—arguably an essential ability for successfully interacting with other agents—may depend on mechanisms other than representational re-coding. Macaques were trained to view photographs of other familiar members of their group. The photographs were either of a mother and her offspring, or of two unrelated group members. After training with the same mother-offspring pairing, the monkeys could successfully judge whether novel combinations were a mother-offspring pair or a pair of unrelated individuals. Macaques displayed a capacity to recognize a social concept such as mother-offspring independent of the physical characteristics of the particular individuals involved. In fact, the pictures of mother-offspring pairs included mothers with infant daughters, mothers with adult daughters, and mothers with sons.

Another study, by Bovet and Washburn (2003), shows that rhesus macaques (*Macaca mulatta*) are able to categorize unfamiliar conspecifics on the basis of their dominance relations. Here, the monkeys were confronting video-clips of agonistic interactions of unknown individuals of their same species. After some observations, the monkeys were able to successfully recognize the dominant monkey in each interaction.

These two examples show that animals such as macaque monkeys, who cannot reason analogically, can nevertheless build up complex social knowledge just from observation and experience with conspecifics. Although it is widely believed that chimpanzees, and great apes in general, have more complex social cognitive capacities than monkeys, the difference is not significant—and it is possible that this belief stems from a bias of researchers of animal cognition and behaviour in favor of chimps rather than from careful empirical investigation (Tomasello and Call 1997, p. 350). The social space navigated by macaque monkeys is of comparable complexity as the social space of chimps notwithstanding their inability for the kind of representational re-coding required by analogical reasoning. Hence, representational re-coding is probably not causally necessary for making possible complex social knowledge and for proficient social navigation. If moral cognition enables the proficient navigation of one's social space that is *not* because of the further capacity to use tags and labels to represent abstract, complex relationships. This conclusion needs some qualification.

There is *no* claim in Clark's argument that chimps have more complex social knowledge or social structures courtesy of their capacity for analogical reasoning. His point with the chimps' case-study is to provide us with a (non-linguistic, non-moral) example in support of the claim that language as an artifact makes possible new cognitive domains.

I agree on the general point that re-coding is a formidable way to make available new cognitive objects which can be used for further thought. Nonetheless, if animals like macaque monkeys who have a limited capacity for representational re-coding present a level of social expertise similar to that of animals like chimps that are more skillful in re-coding, then re-coding is probably not causally necessary for the emergence of a complex social-normative space.

At bottom, moral wisdom might be a type of know-how, enabled by Bayesian-RL neurocomputing, that we share with "baboon troops, wolf packs, dolphin schools, chimpanzee groups, lion prides, and so on" (Churchland 2000, p. 297). In many of these animals we don't witness a capacity for representational re-coding, yet they do display "the same complex ebb and flow of thoughtful sharing, mutual defense, fair competition, familial sacrifice, staunch alliance, minor deception, major treachery, and the occasional outright ostracism that we see displayed in human societies" (Ibid.).

Pattern-recognition and certain types of social learning might suffice for animals like macaques to perform successfully in the tasks described above, and more generally to proficiently navigate their social-normative space (Churchland

2011). Those monkeys had extensive interaction with other conspecifics before confronting the experimental tasks. In the course of such interactions, macaques could learn that certain patterns of cues correspond to certain abstract relationships (e.g. grooming, parenting, sexual interaction, fights or alliances). The identification of different types of relationships need not depend on recollection of all the specific elements instantiated in particular cases. Rather, as suggested by Churchland, it can depend on prototype-based cognitive constructs which deliver the statistical central tendency of a large number of concrete exemplars many of which can differ in important ways from the others. If macaques can extract and keep track of different social prototypes across contexts and act upon them, then they may successfully deal with new social situations by recalling what prototype best corresponds to the particular pattern of cues in that context.

Such library of prototypes, as Chapters 1 and 4 argued, might get imbued with value after patterns of rewards and punishments are received in a given situation (Ibid., Ch.2). For example, assume that a sufficient number of macaques are willing to punish other macaques for behaving in a way *b* which is harmful to the group members. When macaques assist to behaviour *b*, they may have certain prototypes associated to *b* or to aspects of *b*. If some macaques engaging in *b* are punished, then *b* will probably get discounted: macaques will tend to avoid engaging in *b* in certain circumstances. While rewarding objects and events make agents come back for more, negative-valenced objects and events tend to be avoided. Thus some of the prototypes associated to *b* might acquire a negative valence. For example, when macaques that are at a lower level in the social hierarchy eat berries in certain circumstances, other macaques may punish them. Under certain circumstances, that is, eating berries when you are at a lower level in the hierarchy is a behaviour that will tend to be discounted. Thus, the macaques might learn that some types of behaviours, which are represented in certain ways, cause the delivery of punishment, and if they want to avoid punishment, they should behave in certain ways rather than others in certain circumstances.

In this sense some prototypes can acquire value. There is in fact evidence that animals like macaques can "divide the world into distinct in-groups and out groups, associate and categorize novel stimuli associated with these groups, and valence these groups as 'good' or 'bad'—all in the absence of language" (Mahajan et al. 2011, p. 401).

Now, *even if* my argument thus far is sound, my proposal would still face two problems. First, the trial-and-error learning and pattern-recognition capacities on which moral cognition and social navigation might depend are common to most animal species. *But* humans—unlike other animals—do *seem* to display a peculiar norm-hungriness, as Haugeland suggests. If we take this apparent difference seriously, what could account for it?

Second, it seems that my disagreement with Clark is merely terminological. We agree that facts like chimps' capacity for representational re-coding and macaques' social categorization are relevant to explain the emergence and unfolding of moral thinking. Our dispute seems to concern merely the language used to describe this space. Are there substantive points of disagreement between Clark's and my argument?

I tackle these two problems in the next two subsections.

2.1 Human Norm-Hungriness and Florid Control

To begin addressing the first issue, one approach is to ask whether some capacity enabled by some neural innovation could account for our peculiar norm-hungriness as surmised by Haugeland. As the prefrontal cortex is reliably involved in most of what it's taken to be distinctively human forms of thinking and capacities, one promising way to understand our idiosyncratic norm-hungriness is to point to some cognitive capacity enabled by the human prefrontal cortex (on this point see Preuss 2009; Stone 2007). Pursuing this approach, my hypothesis is that humans' peculiar norm-hungriness depends on human capacity for *florid control* which is enabled by the concerted interaction between our prefrontal cortex and the dopaminergic RL-system. I now elaborate my hypothesis.

The prefrontal cortex (PFC), which we already encountered in previous chapters, is the neo-cortical region most complicated in primates. It comprises an ensemble of interconnected areas organized such that they can send and receive neural projections from the sensory and motor systems, and many sub-cortical areas. Compared to other mammals, humans evolved large brains, with a disproportional enlargement of the PFC relative to body size (Kaas and Preuss, 2008). Human PFC differs not just in size but also presents much more morphological complexity (Stone 2007; Preuss 2011). Among the advantages derived from a larger and more complicated PFC, there seem to be sophisticated capacities such as high-level, flexible goal pursuit, planning, selective attention, and working memory (Fuster 2008).

The dopaminergic system in the basal ganglia and brainstem is just behind the PFC. It should be clear at this point that the dopaminergic system is known to play major roles in motor control, learning, motivation and reward-based decision-making (for reviews see Seamans and Durstewitz 2008; Redgrave 2007). Between the PFC and dopamine neurons in the basal ganglia there are strong bidirectional connections which probably indicates that the interaction between PFC and dopaminergic system in the basal ganglia could serve specialized cognitive functions: there is evidence that the PFC could exert top-down regulatory control over the ascending modulatory signals from the brainstem (Robbins and Arnsten 2009), while phasic dopamine signals could attend to the gating of new information to the PFC (Cohen, Braver and Brown 2002). Although the human neo-cortex does not seem to present an increase of dopaminergic innervation in comparison to the neo-cortex of other species, humans present some significant innovations in the morphology of dopamine cortical innervations, which also tells for a specialized role of dopamine in cortical organization occurred in the evolution of the human brain (Raghanti et al. 2008). Neuropharmacological research also shows that PFC functions such as self-control and attention are highly sensitive to changes in dopamine levels (Robbins and Arnsten 2009).

Given the possible computational roles of the PFC-Basal ganglia circuit in the RL-Bayesian mechanism laid out in Chapter 1, and given all the cognitive functions for which the concerted activity of the PFC and the dopaminergic system are necessary, it shouldn't be surprising that they are also crucially implicated in social cognition and moral judgement (Forbes and Grafman 2010). But what could the interaction between PFC and dopaminergic system have to do with human *norm-hungriness*?

The hypothesis on offer is that among the cognitive functions enabled by such an interaction there is what can be called *florid control*. The apparent peculiarity of our norm-hungriness could be accounted for by the human cognitive capacity for florid control. This is the capacity to value and pursue biologically-arbitrary thoughts and behaviours in the face of distracting stimuli, disruptive emotions, competing drives and intentions. Florid control comprises two distinct cognitive capacities: the capacity to value biologically-arbitrary thoughts and behaviours, and the capacity to maintain and follow through thoughts and behaviours while ignoring potential distractive stimuli and suppressing competing and disruptive information. The human dopaminergic system would support mainly the former capacity; the PFC would be essential for the latter.

Biologically-arbitrary beliefs and behaviours are those that do not obviously contribute to life maintenance and reproduction. Chastity and hunger-strike are examples of biologically-arbitrary behaviours. The human dopaminergic system would enable *any* biologically-arbitrary belief and behaviour to be able to gain the status of primary reward like food and water. Read Montague (2007) describes such a capacity as a uniquely human "superpower." What could make this "superpower" possible is a specific pattern of dopaminergic signaling that encourages the rest of our cognitive system to pursue certain beliefs and behaviours while increasing the relative valuation of stimuli that predict them. Phasic dopamine signals might allow such highly-valued beliefs and behaviours to gain access and hold onto the PFC (Montague et al. 2004).

Once such beliefs and behaviours gain this high-value status holding onto the PFC, they can become intrinsically motivating courtesy of the control enabled by prefrontal activity. In some cases these beliefs and behaviours correspond to social norms and normative behaviour. They can lead us to comply with social norms "even when there is little prospect for instrumental gain, future reciprocation or enhanced reputation, and when the chance of being detected for failing to comply with the norm is very small" (Sripada and Stich, 2007, p. 285). *Human* norm-compliers would be able to ignore distracting stimuli and suppress disruptive, competing motivations in the pursuit of the social norms they are hungry for.

Note that florid control needs not be conscious. "A firmly held goal often means that potential distractions are nonconsciously ignored, and that disruptive emotions or drives are nonconsciously suppressed. When social niceties become 'second nature,' one does not have to consciously work out what to do, or consciously suppress intentions that could intrude and make for awkwardness" (Suhler and Churchland 2009, p. 345). Consciousness therefore is not the mark of florid control or of norm-hungriness.

Because other animals lack florid control, they would not display norm-hungriness of the kind displayed by humans. Other animals, unlike humans, cannot be motivated to comply with *any arbitrary* social norm. They do not appear to be able to bestow value onto biologically-arbitrary behavioural patterns. Although they display a capacity to exercise control and select appropriate actions in the pursuit of their goals, their control is not florid. Other animals do not seem to be able to display control over behaviours which do not bring any obvious benefit to their group or to themselves.

In social situations, other animals might pay attention to what others do, and rely on their past experience to learn what behaviour is most appropriate in that situation. Once the appropriate behavioural pattern has been learned, they might act on it. But they cannot act on a pattern of behaviour in spite of the rewards and punishments delivered by others. Dogs, chimps and other non-human mammals have complex social knowledge, can be sensitive to subtle cues in their social environments, they can care for their juveniles, mates, kin and affiliates, can display articulate forms of social interaction, they might even attribute mental states (Churchland 2011; de Waal and Tyack 2003). What non-human animals cannot do is to comply with norms of chastity or hunger-strike.

2.2 Beyond Terminological Disagreement. Local Moral Thought and Moral Systems

Much of my disagreement with Clark seems to hinge on how 'morality' is best defined. I believe that in fact this is not the case: my disagreement with Clark isn't merely a matter of terminology. While I can see at least two points of substantial disagreement, there are also two points where we can be in agreement. Clarifying these points will help us identify ways to make philosophical and scientific progress.

The first point of genuine disagreement is whether language is necessary for agents' *committing* to certain ways of behaving in a community. Moral commitment, for Clark, appears only in moral debate. Moral behaviour, as noted above, would demand a "commitment to finding routes through moral space that accommodate multiple perspective and points of view" (Clark 2000b, pp. 309-10). Language would make such a commitment possible by creating the conditions for agents to enter and solve moral clashes. I disagree with this claim.

Language does not create the conditions for moral commitment. I don't question that moral debate is an important way to display and pursue one's moral commitments in a community. But I'm not convinced that debate and argument are the only ways to make moral commitments and displays of such commitments possible. There are non-linguistic ways that create the conditions for agents to pursue practical agendas in a community. Non-linguistic agents can commit themselves to certain types of courses of action in a community, and display their commitments by relying, for example, on a suite of moral emotions.

Frank's (1988) idea of emotions as *commitment and signaling devices* is relevant here. The idea is that emotions like anger or guilt can commit agents to pursue specific courses of actions even contrary to their immediate material self-interest. Since emotions are typically visible to others and can be hard to fake, they also function as signaling devices. Anger, for example, would signal to others a commitment to aggressive behaviour. Being committed to aggressive behaviour and signaling this commitment could prevent other agents from acting on certain behavioural patterns and thereby clashes could be avoided. If clashes cannot be avoided, being committed to certain emotions can still prompt ways for solving practical issues. Sex or fights are two such ways: they are ways of solving moral clashes divorced from argument and debate. Hence, the emotions may be non-linguistic means to "finding routes through moral space that accommodate multiple perspective and points of view."

The second point of disagreement with Clark concerns the role of language in humans' norm-hungriness. Clark holds that language, via the representational re-coding it enables, is necessary to make available a normative domain of thinking. If, for example, dogs cannot comply with a norm of chastity, that is because patterns of behaviour of certain types are unavailable to dogs' minds. And, in turn, this is because dogs lack language. We can comply with a norm of chastity because we possess language, and language makes our minds sensitive to patterns of behaviour like chastity. For Clark, therefore, language is necessary for our peculiar norm-hungriness. I deny this claim.

If we understand 'chastity' as abstention from all sexual intercourse, creatures with no language do not seem to be necessarily blind to the corresponding behavioural pattern. Non-human animals like macaques or baboons possess complex social knowledge. Macaques, we have seen, can recognize social concepts like mother-offspring. Baboons seem to understand "in what matriline every animal belongs, how the matrilines are ranked relative to each other, and who ranks where within each matriline" (Churchland 2011, p. 127). All this involves a lot of social knowledge which can be acquired courtesy of pattern recognition and social learning, based on other agents' rewards and punishments and on imitation. Baboons' social knowledge facilitates their social decision-making. For example, it facilitates them to act upon specific norms of cooperation, grooming and food sharing.

Given this capacity to acquire rich social knowledge and to act upon it, and given the saliency of a behavioural pattern like chastity, it does not appear wildly implausible that the concept of chastity could be available to the minds of some non-human animals.

Even if such a concept is available to their minds, however, non-humans animals couldn't comply with a norm of chastity because they lack *florid control*. As I suggested above, non-human animals could not value *any* biologically-arbitrary behavioural pattern *and* they could not exercise control resistant to competing drives and distracting information so that they can act on such a pattern. It seems unlikely that dogs, baboons and macaques can act upon norms of chastity, but not because they lack language; they cannot comply with chastity because they lack florid control.

In spite of much disagreement, there are two points where Clark and I can find common grounds. To reach these common grounds we do not need firstly to tackle the question of how 'morality' should be defined. There's room for substantial agreement once we bring Gibbard's (1990) notion of *accepting a norm* to bear on our understanding of 'norm-hungriness,' and we distinguish between *local moral thought* and *systemic moral thought*.

Clark can agree with Churchland and me that other animals, even those with no capacity for representational re-coding, can display moral thought. They can possess knowledge of complex social relationships on which they can act, engage in altruistic behaviour, show empathy, punish norm violators, reconcile after fights, have subtle policies to regulate behaviour in their communities (see e.g. Churchland 2011; de Waal 1996; de Waal and Tyack 2003).

Non-human animals, however, aren't moral in the way humans are:[8] They cannot *accept* a norm in the sense singled out by Gibbard (1990, Ch. 4); and they lack what I call *systemic* moral thought. Clark, Churchland, Haugeland and I can agree that language is probably an essential prerequisite for both *accepting* a norm and *systemic* moral thought. Let me explain.

[8] Churchland (2011, p. 26) writes: 'Of course only humans have *human* morality. But that is not news, simply a tedious tautology. One might as well note that only marmoset have *marmoset* morality, and so on down the line. We can agree that ants are not moral in the way humans are, and that baboon and bonobo social behavior is much closer to our own.'

The aim of Gibbard's (1990) is to understand the nature of rationality and morality in a way that fits with a picture of ourselves as members of an evolved species facing recurrent bargaining situations. In order to reach his aim, the main notion to be explained is that of *accepting a norm*. Gibbard's basic proposal is that "to think something rational is to accept norms that permit it" (Ibid., p. 55).

For Gibbard, accepting norms "is a significant kind of psychological state" unique to humans (Ibid.). He explains: "The state of accepting a norm, in short, is identified by its place in a syndrome of tendencies toward action and avowal—a syndrome produced by the *language-infused* system of coordination peculiar to human beings. The system works through discussion of absent situations, and it allows for the delicate adjustments of coordination that human social life requires" (Ibid., p. 75, emphasis added).

Accepting norms would be the capacity to be motivated to sincerely avow and to act upon certain behaviour patterns which evolved because of the advantages of coordination and planning through language. Successful coordination and planning in bargaining situations faced by complexly social species like ours would require normative discussion. This is the practice of evaluating with one another what to do, think or feel in various, typically absent, situations. Within normative discussion agents tend to be responsive to others' demands and needs, and to be influenced by the avowals of others. The acceptance of norms arises from and is influenced by normative discussion. Since normative discussion is grounded in language, the acceptance of norms seems to depend on and be influenced by language.

Now, Haugeland's notion of norm-hungriness is vague. Above I provided one possible characterization and I argued that norm-hungriness does not depend on language. *If* norm-hungriness is understood in terms of Gibbard's notion of accepting a norm, however, it should be clear that norm-hungriness does depend on language. Norm-hungriness, in *this* sense, would be a capacity that can only be acquired under the pressure of normative discussion, whereby we determine what it is to count as rational, or morally right or permissible, or what we should believe or feel. Norm-hungriness, in this sense, would also lead to *systemic* moral thought.

By 'systemic' I refer to the kinds of effects of moral thought. Non-human animals' moral cognition can have only *local* effects. Courtesy of language, instead, human moral thought can spread throughout space and time by creating cognitive niches that foster a normative explosion (Clark 2006b). Such niches correspond to social structures such as families, churches, governments, markets, legal systems, post offices, hospitals, universities, museums, theatres and so forth. In such niches norms get propagated *and* become themselves objects of moral thinking. Churches, schools and museums not only secure that norms are transmitted by facilitating that people are instructed in the endorsement of evaluative, behavioural and epistemic norms. More importantly, they provide us with the conditions to bring normative considerations to bear on norms themselves. Social structures like churches, schools and museums are niches where normative thought can be objectified thereby promoting a normative explosion where higher-order norms can emerge to manage our endorsement of lower-order norms.

Language, it seems, is an essential prerequisite for the creation and policing of these social structures (Searle 1995). Language would be necessary for the creation of such structures because it provides us with a unique means to collectively represent something as having a certain status beyond its physical features. For example, once a certain building is collectively represented and accepted as having the status of a school, and such collective representation and acceptance is common knowledge, then that building can perform functions that it could not perform before. Language is a unique means to grant a certain status to a something and to make common knowledge its purposes, the legitimate moves with it and within it, the roles attached to different agents engaging with it and within in it.

Language makes possible also to effectively police and give shape to these structures. We can manage and direct such complex structures only because language enables us to reflect on maxims, labels and moral summaries and categories; language enables us to make normative considerations concerning norms themselves. Thus we can manage such social structures so as to facilitate the attainment of determinate effects on the community as a whole, or direct their function towards new purposes.

Therefore, although language probably does not constitute a moral domain of thinking, it dramatically changes its scope. Language is probably constitutive of systemic moral thought as it enables and fosters local effects of moral thought to spread systemically and reiteratively across space and time.

3. Causal Influences of Language on Moral Cognition

Language can have important specific causal consequences on moral cognition and social decision-making. Clark identifies and discusses one such consequence: language would bias selective attention during moral problem-solving (Clark 1996;

2000a, Sec. 3). The remainder of the chapter integrates Clark's discussion by identifying three further consequences that rules, normative maxim, moral discourse, and language more generally can have on moral cognition. Moral instructions can modulate the degree to which rewards and punishments impact social learning; moral labels can trigger looping effects; language bootstraps moral thinking into meta-ethics. I succinctly characterize each in turn, after having critically presented Clark's suggestion about the effects of language on selective attention.

3.1 Language and Selective Attention

Clark argues for a special, although not exclusive, causal role of linguistically encoded norms and summary principles on individual processes of selective attention and decision-making. To illustrate the point Clark (1996, pp. 118-9) recalls Kirsh and Maglio's (1992) analysis of the performance of Tetris players. Kirsh and Maglio argue that expert Tetris players rely both on reactive, pattern-completing cognitive mechanisms, and on linguaform, high-level normative policies which they use to monitor the processes of the former. Normative policies express how things ought to be, that is something the agent should be concerned about. Examples of these policies are "Don't cluster in the center," "Keep the contour flat," "Avoid piece dependencies" (Kirsh and Maglio 1992, pp. 8-9, quoted in Clark, 1996, p. 119). Such policies would bias the processes of selective attention thereby determining the input to reactive, pattern-completing mechanisms. After this bias, Tetris players' behaviour tends to be in line with the policy, and their performance improves.

Clark suggests that the same might apply in the moral domain: explicitly formulated summary rules and moral maxims "may help us monitor the outputs of our online, morally reactive agencies. When such outputs depart from those demanded by such policies, we may be led to focus attention on such aspects of input vectors as might help us bring our outputs back into line" (Clark 1996, p. 119). The idea is that the maxims, laws, normative policies and the linguaform moral rules can bias the workings of our more basic pattern-recognition capacities. Rules would flag cases where current moral practices diverge from the normative ideal, and ultimately they influence our judgements and decisions and lead us to conform to what ought to be the case.

Clark's suggestion needs qualification. Casual observation indicates that much social behaviour is inconsistent with linguistically-codified rules which people are aware of. One reason why this is so is because one's decision to follow a rule in some situation is significantly influenced by what she believes most people do in that situation. When the majority's behaviour in some situation is inconsistent with the rule, people may not expect to be punished if they break that rule. Hence, under the causal pressure of information about what people typically do, selective attention seems to discount what is prescribed or proscribed by some normative policy.

In the case of games like Tetris there are specific standards of success. If we don't follow those standards and don't try to implement some normative policy, the game will be over soon. In the social domain, instead, we don't have specific standards of success. In general, if people don't have a particular personal concern to follow a rule, and the rule flies in the face of typical behaviour of the majority, information collected courtesy of learning and direct observation will be the major causal determinant of their behaviour.

It seems then that laws and linguistically encoded social norms tend to have grips on people's selective attention and decision-making only if they are consistent with information about what most people do (Bicchieri and Xiao 2009). Normative messages that focus on evidence that most people engage in some desirable behaviour are more effective than messages that focus attention on the detrimental consequences of norm violation. For example, in hotel rooms we often find cards asking us to reuse our towels for the sake of helping save the environment or to save resource. But the message communicated by these cards is often ineffective: "Within the statement 'Look at all the people who are doing this *undesirable* thing' lurks the powerful and undercutting message… 'Look at all the people who *are* doing it'" (Cialdini 2003, p. 105). When people's attention is drawn to the fact that the majority of guests do reuse their towels when asked, towel reuse increases significantly (Goldstein et al. 2008). The causal effect of rules and normative policies is often optimized when they are aligned with information about what people typically do.

3.2 Language and Reward-Learning

A second effect that verbal instructions and rules can have on moral cognition concerns learning. Besides trial-and-error learning, verbal instruction is an efficient means to learn how to navigate the social environment. Recent computational and neuroimaging work indicates that verbal information can have significant impact on reward learning (Doll et al. 2009; Li, Delgado and Phelps 2011). When reliable verbal instructions are available, we can assign less weight to observed feedback which can spare us multiple errors and learn more quickly.

Doll et al. (2009) developed two neurocomputational models that could explain the precise effect of verbal information on reward learning: an “override” and a “bias model.” In the first, the striatum—a subcortical brain region and major target of dopaminergic neurons—learns cue-reward probabilities as experienced, but is overridden by the PFC—where instructed information would be encoded—at the level of the decision output. In the bias model action selection and learning supported by the striatum are biased by rules and instructions encoded in the PFC.

These types of models are first attempts to explain the roles and interactions of different types of information affecting learning and decision-making. Yet, it is not clear why linguaform information influences learning and behaviour in some cases and do not in others. Perhaps, verbal instructions and linguaform moral rules have special impact on how people learn from feedback in complex and social situations where basic reinforcement learning may not be the most efficient way for social navigation.

3.3. Language and Looping Effects

Third, normative talk and moral labels can have a *looping effect* on our moral judgement and behaviour (Hacking 1995). Hacking argues that the creation and spread of labels like ‘child abuse,’ ‘multiple personality disorder,’ ‘teen-age pregnancy’ can causally affect the ways we think about and interact with the objects they refer to. The labels and classifications we use to identify a certain human kind influence social behaviour towards the individuals that fall into that category. At the same time such labels shape the self-understanding and behaviour of those that are categorized.

The looping effect ‘is about how a causal understanding, if known by those who are understood, can change their character, can change the kind of person that they are. This can lead to a change in the causal understanding itself’ (Hacking 1995, p. 351). The use of linguistic labels to sort out people can affect what we classify, the classifier and the classifications itself, thereby making possible new ways of self-knowledge.

3.4 Language and Meta-Ethics

Finally, language bootstraps thinking into meta-ethics: an abstract reflection on views, presuppositions and commitments of those who engage in moral debate and practice. Meta-ethics is a species of second-order thinking, which is probably a major consequence of language.

As already noticed, words and sentences can in fact serve as anchors for what Clark terms ‘thinking about thinking’ (Clark 1997, p. 209; Clark 2006b): The capacity to think about our own thoughts, reasons or cognitive profile. ‘To formulate a thought in words (or on paper) is to create an object available to ourselves and to others and, as an object, it is the kind of thing we can have thoughts about’ (Clark 2006b, p. 372).

Linguistic formulations of moral thoughts create the conditions for meta-ethics: ‘creates the stable attendable structure to which subsequent thinkings can attach.’ (Ibid.). Normative statements and moral discourse become ‘anchors’ for reflecting about the meaning, the psychological presuppositions, and the epistemological and metaphysical commitments of our own moral thinking.

Conclusion

The chapter has explored the questions of whether and in which sense moral cognition could depend on language. These questions have been addressed by focusing on Andy Clark's case for a discursive construction of the moral space. It has argued that language is probably not constitutive of the moral thinking and that humans' peculiar norm-hungriness might be underlain by the unique human capacity for florid control. Linguaform normative policies, moral maxims and rules have many distinct effects on moral cognition and on our capacities for moral problem-solving, moral reflection, social learning and decision-making. Four such effects have been identified.

There are a number of important questions I have overlooked. For example, how can language contribute to the persistence of certain norms? How can certain uses of language induce pro-social behaviour? Do language disorders impair the capacity to navigate the moral space? More generally, does it make sense to try and identify the aspects, if any, of our moral practice that fundamentally distinguish us from other animals? And what type of empirical evidence can bear on such an issue?

CHAPTER 6.

Caring, Emotions and Social Norm Compliance

This thesis has argued that the mechanism of norm compliance probably consists of RL-Bayesian neurocomputations. It has claimed that people in complying with norms are subject to many sources of motivation, and that social representations, by themselves, are not sufficient to motivate[9] norm compliance. The reward-values attached to social representations courtesy of RL-systems are also necessary. Now, after having claimed in the previous chapter that humans' peculiar norm-hungriness might depend on *florid control*, I want to examine more closely some of the aspects of the motivational structure of norm compliance, at both the personal and subpersonal level.

Emotion is the focus of this chapter since there is little doubt that it plays a crucial role in the regulation of our moral and social life. Yet, it is controversial in what sense emotion motivates people to abide by social norms. The empirical evidence doesn't warrant firm conclusions and the philosophical debate has mainly focused either on emotion and norm violation, or on the relationship between emotion and normative judgement (see e.g. Sinnott-Armstrong 2008). This chapter asks three questions relevant to understanding the personal and subpersonal natures of the reward-values computed by the RL-system:

1) Are emotions or emotional processes generally the ultimate motivational source of social norm compliance?

2) Are the reward-values computed by RL-algorithms in the striatum best understood as emotions?

[9] With 'motivation' I refer to processes that influence the triggering or direction of norm compliance behaviour. 'Ultimate motives' (sometimes also called 'primary motives') are the starting points of causal chains that lead to action.

3) How could we characterize the capacity to care, on which the motivation to comply with social norms seems to depend, within the neurocomputational framework put forward in the previous chapters?

The answers I shall argue for are:

1a) The emotions are not the ultimate motivational source of norm compliance. The capacity to care is probably necessary both to feel emotions and to comply with social norms.

2a) There is little evidence that the reward-values computed in the striatum should be understood as emotions—as hedonic units in particular.

3a) One way to give neurocomputational flesh to the capacity to care is in terms of the computational dynamics of various neuromodulatory systems.

The chapter is in three sections. Section 1 tackles the first question by engaging with one of the few explicit arguments that the emotions are the ultimate source of norm compliance: Robert Sugden's *Resentment Hypothesis* (Sugden 1998; 2000). Sugden's argument is congenial to this chapter—which does not hinge on any sophisticated account of emotions—because it seems to assume a commonsensical view of emotions understood as feelings that people experience. I argue, *contra* Sugden, that the emotions—in this sense at least—are not the motivational source of norm compliance.

With the results of my critique to Sugden's account in hand, I tackle the second and the third question. Section 2 argues that Fehr and Camerer's (2007) hedonistic interpretation of neurobiological data about social norm compliance is unjustified. The reward-values computed by the brain mechanisms that might implement RL algorithms should not be understood in terms of pleasure. The last section suggests that caring, which is probably bound up with social norm compliance, might depend on a fundamental aspect of the RL-system. What we care about might be determined by the setting and adjustment of several parameters in RL-algorithms courtesy of specific neuromodulatory systems.

1. Emotion and Norm Compliance

Imagine you are travelling on a crowded train without a seat. While you are tired of standing, someone leaves her seat to go to the toilet. Why don't you take her seat? A plausible explanation may invoke the existence of a norm that bounds the set of appropriate actions in that type of context. In the vocabulary of folk-psychology: because you believe that taking the seat of someone who leaves it to go to the toilet falls outside that set and you find that norm reasonable, you don't take the seat and you keep on standing.

Robert Sugden (1998; 2000) argues that it is *not* your acceptance of the norm that plays a fundamental role in motivating you to comply with it. Sugden develops an "emotional sanctioning" account of norm-compliance. One of the aims of his work is to explain where the "feeling of normativity" comes from. He aims to explain the emergence of social norms in general, and norm compliance in particular, with no appeal to normative concepts.

Sugden's argument is in two stages (Sugden 2000, Sections 3-4). The first leads to the formulation of an empirical hypothesis called the *Resentment Hypothesis*, which provides us with sufficient conditions for the arousal of resentment. Resentment, for Sugden, is a non-moral sentiment which does not depend on any moral code. What ultimately motivates norm compliance would be the sensation of resentment. The second stage in his argument aims to defend the psychological plausibility of the Resentment Hypothesis.

Before examining Sugden's Resentment Hypothesis, I succinctly clarify how 'emotion' is used here. 'Emotion' is a contentious term. Sugden uses 'emotion' interchangeably with 'sensation,' 'sentiment,' 'affect,' and 'feeling.' He seems to be influenced by Adam Smith's (1759/1976) theory of moral sentiments (see also Sugden 2002 on this point). Smith in fact provides a commonsensical account of various feelings such as resentment and sympathy, which we are invited to test against our own experience. Accordingly, I use 'emotion' in an ordinary sense, as a type of feeling (see Bennett and Hacker 2003, Ch. 7; for accounts that deny that emotions are types of feelings see e.g. de Sousa 2010).

In this sense, emotions are mental episodes that one experiences, and their essential feature is their qualitative character. This ordinary notion of 'emotion' has two distinct aspects, which will be important in relation to Sugden's argument. 'Emotion' can refer both to emotional perturbations and emotional attitudes. Emotional perturbations are episodic, short-lived states. Some emotional perturbations, such as outbursts of anger, are accompanied by characteristic somatic changes, which can include increased heart-beat rate, sweating, muscular tension and throbbing temples. Other emotional perturbations, such as feelings of pride, manifest in expressive behaviours such as when one issues utterances of pride, changes in posture or in tone of voice.

Emotional attitudes last for longer periods. Love and hate, guilt and regret are emotional attitudes that can last for years. For example, love as a standing attitude of fraternal feeling is distinct from love as the episodic perturbation of falling in love with a boy. Love as an emotional attitude is persistent; it can motivate certain kinds of actions and thoughts towards the beloved even after the initial perturbation has gone. Both emotional perturbations and emotional attitudes often motivate people to comply with social norms. For example, you may refrain from taking somebody else's seat on a crowded train because of a negative emotional pang, or because of a long-standing shame of misbehaving in social situations like that. Having clarified my usage of 'emotion,' I now examine Sugden's Resentment Hypothesis on the relationship between emotion and norm compliance.

1.1 Bob Sugden's Resentment Hypothesis

Sugden begins by claiming that when other people's actions constitute a predictable behavioural pattern, they thereby *seem* to impose "some obligation on me to conform to that pattern" (Sugden 2000, p. 112). The claim is not about the existence of a general moral principle. It is not that there exists some obligation to conform to behavioural patterns in virtue of their being predictable. The claim is that "people are *in fact* motivated as if by some such principle" (Ibid.). Certain behavioural patterns are associated with particular *normative expectations*. And normative expectations motivate us to comply with norms courtesy of specific affective signatures. When one has a normative expectation, she expects that others expect her to do something. But how is it that the fact that some people expect one to do Φ in a certain type of situation S makes her want to do Φ in S?

According to Sugden, we naturally feel resentment against those who act contrary to our expectations and we also feel aversion towards frustrating others' expectations. By 'resentment' Sugden means "a sensation or sentiment which compounds disappointment at the frustration of one's expectations with anger and hostility directed at the person who is frustrating (or has frustrated) them" (Ibid., p. 113). Aversion depends on resentment. One conforms to a behavioural pattern because others will resent her otherwise, she knows this, and she is emotionally averse to others' resentment. In many situations resentment and aversion are intertwined with cognitions. 'Cognition,' recall, here refers to processes supporting such mental states as knowledge or belief that contrast with affective or emotional processes.

There are two ways in which cognitions enter Sugden's account of norm compliance. First, he acknowledges that sometimes people feel resentment and they have knowledge that they have been wronged, given some normative standard. But people do *not* feel resentment *because* of their normative knowledge. That person j feels resentment at person i's doing Φ doesn't presuppose that j believes that i ought not to Φ. For example, your friend and you have agreed to meet for lunch. You are waiting for her, when she phones you telling you that she is ill and she cannot make it. Although you know that your feeling is unjustified, you may feel resentment towards your friend in this situation. Similarly, that person i feels aversion towards doing Φ doesn't presuppose any belief by j that he ought not to Φ.

For Sugden, resentment and aversion are more fundamental than ought-beliefs in two ways. On the one hand, resentment and aversion as sensations are evolutionarily more primitive than cognitions, such as ought-beliefs. On the other, many of our ought-beliefs "are nothing more than generalizations of more primitive sentiments" like resentment and aversion (Ibid., p. 115). When ought-beliefs have the power to motivate people to comply with norms in particular cases, they do so in virtue of resentment and aversion of which they are generalizations. Hence, resentment and aversion are also more fundamental than ought-beliefs in motivating norm compliance in particular cases.

There is a second way in which cognitions may be linked to resentment and aversion. This leads us to the formulation of the *Resentment Hypothesis*, which relies on common knowledge conditions. Specifically:

> Let P be a population and I a behavioural pattern dependent on some interaction among the individuals in P. Let i and j be any two individuals from P that engage in I. Let Φ and Ψ alternative actions that i can take in situation S. Whichever action i decides to take, it will be common knowledge after the event. Assume that it is common knowledge within P that individuals in i's position normally do Φ rather than Ψ. It is also common knowledge within P that people in j's position have grounds to expect i to Φ and that they normally prefer that people's in i's position do Φ rather than Ψ. Granted that j has that preference, then i's doing Ψ will induce in j a feeling of resentment towards i; and i's being aware of this will induce in i a feeling of aversion towards doing Ψ (Sugden 2000, pp. 114-116).

Sugden's Resentment Hypothesis says that people will feel resentment towards those who fail to conform to their expectations. Because this tendency of people feeling resentment is common knowledge, people will tend to avoid acting in ways so as to provoke feelings of resentment. The hypothesis is stated as sufficient condition for the arousal of resentment. The bottom line is that a "person can be motivated to meet other people’s expectations about him” and this motivation is grounded in an emotion (Ibid.).

Sugden illustrates how the sentiment of resentment explains norm compliance with the following type of example. It is well-known that diners in the United States leave tips of at least 15% of the bill. I know this fact. I have good reason to expect that waitresses in the United States expect me to leave a 15% tip if I dine out in the US. I go to a restaurant in the US, but I am Italian and it’s not in my interest to meet the waitress’ expectation. Still, the existence of the expectation will motivate me to tip her. If I don’t tip, I will feel uneasy and embarrassed. I am emotionally averse to those emotions, and this aversion motivates me to comply with the norm of tipping.

1.2 Not by Resentment Alone

I believe that Sugden’s hypothesis is not sufficient. My claim is that the Resentment Hypothesis seems plausible only within a population where people *care* for each others’ preferences, expectations and behaviour. The notion of caring I have in mind will be articulated firstly by ostension, by pointing to the relevant phenomenon with a number of cases. Then, in the following subsection, I shall attempt to elucidate what ‘caring’ means here more carefully.

The argument developed in this section can be summarized thus:

> P1. Sugden’s Resentment Hypothesis depends on an individual j preferring agent i doing Φ.

P2. If an individual j feels resentment about agent i doing Φ then j cares about i doing Φ.

P3. Caring is distinct from preferring.

P4. Sometimes an individual j prefers things about which she doesn't care about.

P4'. Sometimes j prefers i to do Φ while j doesn't care about i doing Φ.

C1. Sometimes j doesn't feel resentment that i doesn't do Φ even if j prefers i to do Φ.

C2. Sugden's Resentment Hypothesis is in general insufficient.

P1 describes one of the conditions in the Resentment Hypothesis. P2 claims that caring about something is necessary for feeling emotions about it. More precisely, we should distinguish between two issues: under what conditions we feel resentment, and under what conditions we are affected by other people's resentment towards us. P2 can be understood as making two claims: we feel emotions only for people, objects, behavioural patterns we care about; we are emotionally affected by other people's resentment towards us only if we care about what other people feel, prefer or think about us. P3 and P4 are related. P4' is a special case of P4. C1 and C2 follow from the five premises. I start by focusing on P2.

Elizabeth Anderson's (2000) can help motivate such claims. Anderson argues that Sugden's account is incoherent. She focuses on the conditions under which people's decision to comply with norms is affected by others' resentment towards them. Sugden—she reasons—assumes that people can feel resentment on behalf of others since we all share the same basic non-moral sentiments. But then norm violators should resent themselves: They need not be averse to others' resentment to be motivated to comply with norms. "Given the impartiality of moral sentiments, they can just as easily be directed against [themselves] as against any other person" (Ibid., p. 184). Hence, other people's normative expectations could be superfluous in motivating one to comply with norms. If self-resentment can be enough for norm compliance, then one can care about complying (or not complying) with social norms independently of what others expect her to do. In other words, people can have an *intrinsic motivation* to comply with norms: they can "comply with norms as *ultimate ends*, rather than as a means to other ends" (Sripada and Stich 2006 p. 281).

A criticism to Anderson's argument is that in general the motivating power of normative expectations, or others' resentment, is greater than self-resentment. Others' resentment causes embarrassment and shame in the violator. These emotional sanctions work as norm-enforcers, and it is the aversion or fear towards such emotions, rather than some sort of intrinsic motivation, that generally motivates norm compliance. If aversion or fear of others' resentment—as opposed to self-resentment or other sorts of intrinsic motivations—has generally more grips on norm compliance behaviour, then people will tend to be less norm-compliant or behave much less pro-socially in anonymous or private conditions compared to what they do publicly. There is in fact experimental evidence that when their choices cannot be detected by other players, participants of economic games tend to behave more selfishly, or so as to merely *appear* to be fair without *being* fair (Bicchieri and Chavez 2010; Dana et al. 2007). Norm-abidance and pro-social behaviour would then depend more on what other people expect from the decision-maker than on some intrinsic motivation.

But there are two problems with this criticism. First, a large number of studies in experimental economics also show that people are often motivated to repay gifts and punish violations of certain social norms in anonymous, one-shot interactions with genetically unrelated strangers, even at substantial costs to themselves (Fehr et al. 2002; Gintis et al. 2003). Even in games with asymmetric information like Dana et al.'s (2007), a significant proportion of participants behave pro-socially both in public and in private conditions. This body of evidence indicates that in experimental situations people generally behave pro-socially or comply with norms not only because they are averse to others' resentment, but also out of intrinsic motives. In real-life situations, depending on the cues and the information available in a given context at a given time, aversion or fear of others' resentment can have more *or less* motivational grip than self-resentment or other sorts of intrinsic motives (Cialdini and Goldstein 2004). People can therefore care about complying with social norms independently of what others expect them to do.

Secondly, conceptually, it seems that people should already care about others' normative expectations in order for those emotions to have some grip on their minds. If I don't care about others' expectations, preferences and behaviour in a certain situation, then I shall probably be indifferent to their resentment. Along these lines, Anderson concludes "[emotional] sanctions are only a supplementary motive to the original motive for compliance, without which the norm would never have been established" (Ibid., p. 184). What I wish to emphasize here with Anderson's argument is that, psychologically, the motivational source of compliance appears to reside in *the capacity to care*. To illustrate and give grounds for this point I now provide some counterexamples to Sugden's Resentment Hypothesis aiming to show that its conditions are insufficient for the arousal of resentment. The bottom line is that caring for other people's preferences, expectations and behaviour is a necessary condition for the arousal of resentment—and for norm compliance.

Consider this situation. After their weekly reading group the participants regularly go to the pub. Ana Maria and Angelica are two of the reading group goers who normally go to the pub. "Going to the pub" and "Not-going to the pub" are alternative actions open to Ana Maria in that type of situation. Within the reading group goers it is common knowledge that a person in Ana Maria's position normally goes to the pub rather than not. It is common knowledge that Angelica has good grounds to expect that Ana Maria will go to the pub. It is also common knowledge that people in Angelica's position prefer that people in Ana Maria's position go to the pub rather than not. Would this be sufficient for Angelica to feel resentment if Ana Maria doesn't go to the pub today after their reading group?

I don't think so. Sugden's Resentment Hypothesis is fulfilled, yet this fails to qualify as a case where resentment is aroused. Ana Maria and Angelica are not close friends; Angelica might be surprised or curious for why Ana Maria is not going to the pub, but she hardly will resent her. To explain why I don't think Angelica would resent Ana Maria, consider another situation.

It's Kirsty's birthday and Rhiannon is Kirsty's best friend. Kirsty has invited Rhiannon to her birthday party. "Going to the party" and "Not-going to the party" are alternative actions open to Rhiannon. Now, would the Resentment Hypothesis be sufficient for the arousal of resentment in Kirsty if Rhiannon doesn't go to her party? It is reasonable to believe that in this case Kirsty would feel resentment. In contrast to the situation above, now Kirsty and Rhiannon are friends and they care for each other. To better illustrate the relevant phenomenon of caring, I point to yet another example.

I live in the Edinburgh area. I read in the newspaper that Miss Carr was found driving on the wrong side of the road in Leith, which is part of the Edinburgh area. I don't know anyone in Leith, I have never been there, and don't plan to go there. Is it plausible that I would feel resentment, in Sugden's sense, towards Miss Carr? Again, I think it is not. In this case, both Miss Carr and I are part of the general population P of drivers in the Edinburgh area. P is quite large. The conditions in Sugden's Resentment Hypothesis are fulfilled, yet it would be implausible to think that I will feel "a sensation or sentiment which compounds disappointment at the frustration of one's expectations with anger and hostility directed at the person who is frustrating (or has frustrated) them." I won't resent Miss Carr even though she frustrates my expectations in this situation and I may interact with her in the future because her behaviour does not matter to me.

This last example also illustrates that in *real-life* situations, when we deal with less close people, we tend to care less about their preferences, expectations and behaviour. Such people are typically members of other groups, so they are not close to us in a literal sense as well: both spatially and temporally (on ingroup-outgroup and social preference see Bernard et al. 2006; Chen and Xin Li 2009). In *real-life* situations, especially when a population is large and it is unlikely that one individual will come to know and interact personally with another individual j, i will not tend to resent actions by j that frustrate her expectations. Also, in general, i will not tend to resent actions by j that frustrate her and another individual k's expectations if it is unlikely that i will come to know and interact personally with either j or k. In real-life, that is, we seem to care more for people we are close to, people we regard as important to ourselves. Note that this claim is consistent with the experimental evidence mentioned above that participants often have an intrinsic motivation to comply with social norms. And in fact in experimental settings that more closely resemble everyday life participants tend to behave more generously with closer individuals (Hoffman et al. 1996; Charness and Gneezy 2008).

If the analysis of these cases is roughly correct, then Sugden's hypothesis is probably insufficient for the arousal of resentment. Feelings of resentment arise in an individual not just because her expectations are disappointed. People seem to feel emotions only about things that matter to them, things they care about. If people feel no emotion about things which they don't care about, then they will feel resented when their expectations are disappointed only if they care about the object of those expectations. If feeling resentment and aversion of being the focus of others' resentment depend on caring, then the Resentment Hypothesis is not sufficient to explain in general norm abiding behaviour.

1.2.1 Caring and Preferring

Here is an objection to the claim that Sugden's Resentment Hypothesis is insufficient because people feel emotions only about things they care about: 'Preference' can be considered a free parameter in Sugden's account and can take different strengths. Caring about something would amount to having a strong preference for that something, and so P3 would be false—and P4 and P4' would be incoherent. Therefore my argument would be entirely consistent with Sugden's Resentment Hypothesis.

This objection is problematic however. To begin with, *even if* we agree that preferences have different strengths and that 'care' can be treated as 'strong preference,' nothing in Sugden's formulation of the Resentment Hypothesis suggests how to identify an adequate threshold for the preference parameter. An individual j may prefer that people in i position do Φ rather than Ψ. Still these pairwise preferences (for i doing Φ over i doing Ψ) might remain below a certain threshold. For example, if the strength of a preference is measured on an interval from 0 to 1, the preference of j for i doing Φ can be 0.2 while j's preference for i doing Ψ can be 0.1. Sugden's conditions are satisfied, but if the preferences are so weak, it seems implausible to think that the Resentment Hypothesis is sufficient to raise resentment in j when i does Ψ instead of Φ. A further condition is required in Sugden's formulation that specifies a suitable threshold such that i's preferences, expectations and decisions do matter to j.

Yet, it may be protested that we could empirically uncover the value of the preference parameter such that if one's preference is unsatisfied she will feel resentment. Different people may care more or less about the expectations and beliefs of others, or perhaps in different situations we care more than in others. By examining possible correlations between choice behaviour and non-choice data like emotional reactions in a given context, a threshold for the preference parameter for resentment arousal could be identified. In this sense, Sugden's account is sufficient as it stands.

But in this sense, the concepts of *what one cares about* and *what one prefers* are assumed to be identical. If preferring is not the same as caring—as P3 above claims—then there are grounds to argue that in some sense Sugden's Resentment

Hypothesis is insufficient. For—as claimed by P4 and P4'—it might be the case that, in some sense, j's preference about i's behaviour and expectations are strong but those expectations and behaviour don't really matter to j: j doesn't care about them. And if one does not feel resentment about something unless it matters to her, then j won't feel resentment when her preferences are frustrated.

There are two questions then: First, what does it mean to care about something? Second, what is the relationship between caring and preferring? Would it make sense to say that an individual (strongly) prefers A over B and yet she doesn't really care about A? My answers to these questions heavily rely on Harry Frankfurt's (1982; 2004) analysis of caring. Let's start from the latter question.

To care about something is not simply to prefer, desire or want it. Attributing a preference "to a person does not in itself convey that the person cares about the object" she prefers over another (Frankfurt 2004, p. 11). Many of our preferences and desires are "utterly inconsequential. We don't really care about those desires. Satisfying them is of no importance to us whatever" (Ibid.). For example, in this moment I prefer to drink water over coke. As I am drinking coke, my preference is unsatisfied. But I don't feel any frustration since I don't really care about such a preference. Note, however, that my drinking coke now does make some difference to me, as everything does make *some* difference to us. This suggests that things we deem important to us, and hence things we care about, are not simply things that make *some* difference to us. Having coke and not water right now is a difference unimportant to me: it's a difference that does not make difference to me. As argued by Frankfurt, “nothing is important unless the difference it makes is an important one” (Frankfurt 1982, p. 259).[10]

This lack of caring and frustration need not be because my preference is weak, or has low intensity. “Sheer intensity […] implies nothing as to whether we really care about what we want.” Frankfurt goes on to explain: “Differences in strengths of desires […] may be radically incommensurate with the relative importance to us of the desired objects” (Ibid.). In the case of preference, from the higher strength of my preference for reading a book over doing the laundry, it does not follow that I especially care about the object of this preference. Even if I intensely prefer one over the other, the difference that reading a book instead of doing the laundry makes to me is not especially important to me now.

Furthermore, “a person who wants one thing more than another may not regard the former as being any more important to him than the latter” (Ibid., p. 12). Frankfurt makes this claim stick with an example. Suppose that you need to kill time and you decide to watch the television. You start to watch a certain program because you prefer it to the others that are available. “We cannot legitimately conclude that watching this program is something that [you] care about.” After all you are killing time. “The fact that you prefer it to the others does not entail that you care more about watching it than about watching them, because it does not entail that you care about watching it at all” (Ibid.). By the same argument, the fact that the individual i prefers that j does Φ rather than Ψ does not entail that i cares more, or at all, about j doing Φ than j doing Ψ.

[10] It should be noted with Frankfurt that “whether a useful account of the concept can be developed without running into this circularity is unclear” (Frankfurt 1982, p. 259).

Suggesting that preferring and caring are distinct concepts is also the empirical finding that our preferences are subject to powerful contextual influences (Lichtenstein and Slovic 2006). There may not be stable facts about one's preferences independent of the way a given choice situation is framed. Caring, understood in a way to be made clearer in a moment, is more stable. "A person can care about something over some more or less extended period of time. It is possible to desire something, or to think it valuable only for a moment. […] But the notion of caring implies a certain consistency or steadiness of behaviour; and this presupposes some degree of persistence" (Frankfurt 1982, p. 261).

If caring and preferring are distinct, what does it mean to care about something? To care about something is not simply to desire it, or want it, or prefer it over something else. Caring is not the same as factoring in things that make *some* difference. In general "caring about something may be a complex mode of wanting it" (Frankfurt 2004, p. 11). For Frankfurt, the capacity to care about something can be understood more precisely as the capacity to commit ourselves to our own desires, wants and preferences. Caring, that is, is a mode of the will.

When people care about something, according to Frankfurt, they desire to have a desire for it, and they endorse such a desire. If a person cares about something, then she is willingly committed to her desire about that thing: she desires that she desires it (Ibid., p. 16). Thus, Frankfurt explains: "by its very nature, caring manifests and depends upon our distinctive capacity to have thoughts, desires, and attitudes that are *about* our own attitudes, desires, and thoughts" (Ibid., p. 17).

Note that this does not mean that all we fundamentally care about is ourselves and our well-being. It does *not* mean that we care about other people's preferences, expectations, and behaviour only because we care about our welfare and well-being. I can care about a waitress expecting me to leave a tip after my dinner *and* comply with a norm of tipping, even if I am aware that by meeting her expectation I won't feel or be better off—next chapter will provide some experimental results relevant to this claim.

In sum, according to Frankfurt, "these alternative possibilities—commitment to one's own desires or an absence of commitment to them—define the difference between caring and not caring" (Ibid., p. 21). It should be clear that, as Frankfurt characterizes it, caring about something is peculiar to members of our species since it requires the ability to *reflexively* deal with higher-order desires. This ability, it appears, is related to what I called *florid control* and to Gibbard's *accepting a norm*, which we encountered in the previous chapter. These three abilities—it seems—are grounded on reflexive thinking on the one hand, and on the ability to commit oneself resiliently to distinct courses of actions on the other. So, strictly speaking, non-human animals cannot care in Frankfurt's sense.

Yet, we can understand caring more broadly than Frankfurt so that we can make room for the possibility of non-human animals that care. Fisher and Tronto (1990) offer a broader characterization, according to which caring is "*a species of activity that includes everything we do to maintain, contain, and repair our 'world' so that we can live in it as well as possible.* That world includes our bodies, ourselves and our environment, all of which we seek to interweave in a complex, life-sustaining web" (Fisher and Tronto 1990, p. 40).

This definition is in line with Frankfurt's: it construes caring as a complex activity supported and informed by a commitment to those desires and goals we deem important for us and for our lives. According to this definition, however, commitment to one's own desires need not be reflexive nor involve self-awareness, or florid control. Agents can hold on to their desires in a persistent, steady way without being conscious of their commitment. Caring, in this sense, would correspond to a relatively stable volitional profile, something with reference to which agents steadily orient themselves in their behaviour and in their environment in the pursuit of a good life.

Also non-human animals, in this sense, would have the capacity to care. And, as a matter of fact, also non-human animals care about staying alive, about avoiding injuries, predators, hunger, thirst and disorder; they may care about close kins, friends and other members of their group; and some may even care for strangers under certain circumstances (Churchland 2011, Ch. 3). So, caring can both be self- and other-directed. And both humans and some non-human animals can care not only about self, but also about others.

It is not obvious what mechanism could ground the capacity to care. Patricia Churchland has recently suggested that hormones such as oxytocin and vasopressin, which originally evolved to promote self-preservation and care for offspring, probably constitute basic features of the mechanism for caring. In their evolutionary trajectory, these hormones would have later been co-opted to serve new jobs so as to enable wider forms of sociability and ultimately to foster moral cognition.

The last section of this chapter integrates Churchland's proposal by pointing to some neurocomputational features of a putative mechanism for caring within the RL-system elaborated in Chapters 1 and 4.

1.3 Evolutionary Origins of the Resentment Hypothesis

The target of the second stage of Sugden's argument is the objection that the Resentment Hypothesis is not reducible to psychology because it is loaded with social and cultural content. To counter this objection Sugden considers the possible evolutionary origin of normative expectations. He asks us to consider an environment akin to a mixed-motive game such as Chicken which is assumed to stand for the environment of evolutionary adaptedness of our ancestors' neurocognitive mechanisms. In such an environment individuals of the same species have to compete repeatedly and enter conflicts for fitness-enhancing resources that are scarce. An adaptive strategy in this game is to act aggressively with weak opponents, and to back down with aggressive stronger players.

The abilities that would enable agents to pursue this type of strategy are three according to Sugden. Firstly, agents should recognize and project patterns in the behaviour of others. A capacity for pattern-recognition would enable agents to identify behavioural patterns of different types of other agents. This would be a prerequisite for behaving in function of the situation and the agents that one is facing. Second, agents should desire to act aggressively against weak individuals. Finally, agents should be averse to acting aggressively against angry individuals. Endowed with these abilities agents can behave so as to get as many resources as possible. Agents can identify that some opponent is trying to frustrate their desire for the resource; but in that type of situation, against that type of opponent, those agents normally get the resource. They can then act aggressively at their opponent, thereby raising their probability of obtaining the resources. Since anger and fear—Sugden (2000, p. 118) reasons—are intrinsic components of the second and third ability respectively, anyone acting on these emotions behaves adaptively. Hence, in a world akin to a game of chicken, anger and fear are adaptive. But what do anger and fear have to do with resentment?

Recall that Sugden defines resentment as “a sensation or sentiment which compounds disappointment at the frustration of one’s expectations with anger and hostility directed at the person who is frustrating (or has frustrated) them” (Ibid., p. 113). Resentment is different from anger. As Sugden acknowledges, resentment is typically backward looking since it is characteristically experienced when one is “looking backward to past injuries” (Ibid., p. 118). How could resentment be evolutionary adaptive?

Sugden draws on Frank’s (1988) account of the emotions, which was introduced in the previous chapter, and argues that anger is a commitment and signaling device. Anger predisposes the angry agent to act aggressively in her next interaction with another agent. The angry individual could incur an immediate cost, but may derive greater benefit in the long run by deterring future frustrations of her desires or injuries. Angry agents would be more likely to get away with some resource for which they are competing with others. This advantage leverages anger as functioning as a signaling device. Other agents need reliably identify angry agents so that they have the opportunity to avoid them. By functioning as a signaling device, anger provides other agents with information about the state of the individual they are confronting. Thus, they will be in the best position to identify agents committed to aggressive behaviour. Put differently, by signaling their commitment to aggressive behaviour, angry agents will be more likely to be avoided by other agents, and consequently to attain some resource without having to fight for it. Resentment, Sugden suggests, may be the side-effect of the evolutionary advantage attached to anger as a commitment device because commitment to aggression could depend on looking back at past injuries. “Backward-looking resentment—Sugden concludes—is the evolutionary price that has to be paid for the advantage[s] of anger” (Ibid, p. 119).

Sugden recognizes that he’s telling us an evolutionary tale about how resentment may be a basic feature of human psychology. His aim is to argue for the *possibility* of reducing the Resentment Hypothesis to psychological features thereby countering the objection that it cannot be reduced since it has too much cultural and social content. I argue, however, that we have more reason to think that resentment, as understood by Sugden, is not a basic feature of our psychological make-up. It is unclear that anger is an adapted feature of human psychology, and it is controversial, at best, that resentment is a by-product of anger. In the remainder of this section I first argue that anger might not be an adaptation, and then I question the link Sugden draws between anger and resentment.

Sugden’s argument is endorsed on more or less the same grounds by evolutionary psychologists such as Tooby and Cosmides (2008, p. 131-132). They argue that anger is an adaptation, which was selected in response to survival challenges faced by our Pleistocene ancestors. Some arguments put forward by some evolutionary psychologists are often charged with mistaking explanation for evidence for the *explanandum* itself (Griffiths 1997). To carry weight, evolutionary explanations of psychological traits should be backed by independent evidence since “adaptive hypotheses are too easy to form and too difficult to test” (Griffiths 1997, p. 71). In the case of emotions, they should be supported at least by evidence about their purported mechanisms and actual functional significance (Machery Forthcoming).

I agree with Sugden that in a mixed-motive game like Chicken it can pay off to act aggressively with anger. It could be true that angry agents often get away with resources, and thereby they have evolutionary advantage over competitors. However, it also could be that they incur long-term costs—especially if chicken was not one of the games our ancestors played more frequently. On the one hand, anger prepares people to overcome obstacles to goal attainment. On the other, however, display of anger leads others to not deal with angry agents (Marsh et al. 2005). Elster (1998, p. 72) argues that angry people will probably "gain more in each interaction, but interact more rarely." Hence, they will not receive resources and feedback which are only available from cooperation with others. The effects of anger, therefore, may be negative overall. "One cannot show that [the effect of anger is] positive simply by citing a positive impact in isolation from other effects" (Ibid.).

Sugden might object to this conclusion that his goal is more modest than we have assumed it to be. He aims to provide something like an existence proof for the evolutionary origin of anger. This doesn't entail an adaptationist approach towards anger. It entails that anger is a universal emotion and it emerges very early in infancy. That anger is universal means that people in all cultures have a similar emotional reaction to things that offend them. That anger appears early in infancy means that culture and socialization don't make a crucial contribution to its emergence.

The problem is that there is evidence that supports the hypothesis that anger may be an emotion differentiated from a generalized, more basic, evolved negative emotion. The innate correspondence between facial expressions and the emotions—which has been traditionally taken as evidence for the evolutionary origin of the emotions since Darwin (Ekman and Friesen 1971)—has been challenged from a developmental perspective. Camras (1992), for example, presents data from a study of infants' early expressive development that show that sadness, anger, discomfort, pain are characteristically displayed together across situations such as being bathed, having a pacifier taken away, exposition to unusual masks, and so forth. Infants often appear to display facial expressions customarily caused by distinct situations within the same burst of cry. These findings give us ground to conclude that anger may emerge as discrete emotion from a basic undifferentiated state of distress through a complex process of socialization (Lemerise and Dodge 2008).

Let us now consider the alleged universality of anger. Prinz (2004, p. 151) reviews anthropological evidence that indicate that some populations lack a word for anger and that people in different cultures respond in different ways to things that annoy them. If the language used to describe and express anger varies across cultures and times in history, and if angry reactions to annoying or offending things are significantly different, then culture might play an important role in the construction of angry reactions. Anger itself might not be a single universal emotion. From this type of evidence, Prinz (2004, p. 151) concludes that although anger is extremely likely to emerge, it is not inevitable. Anger not only regulates social interaction, but is itself partly constituted and comes to be regulated by social dynamics. Therefore, there's reason to doubt that anger is an adapted basic emotion as required by Sugden's argument.

Moreover, if culture and socialization are necessary for anger to emerge, then it seems that social norms should be already in place to give rise to anger-displays, and hence to backward looking resentment. So, in a sense, Sugden's evolutionary argument about the origins of anger and resentment might depend on the pre-existence of social norms. Rather than being explained by the emotions, social norm compliance would explain the emergence of certain emotions.

What about the link between anger and resentment? Sugden claims that resentment would be one of the effects of the evolutionary advantages reaped by angry agents since being committed to aggressive behaviour would depend, to a great extent, on looking back with anger to past injuries. There are two reasons to think that this claim is unjustified. The first draws on the distinction made above between emotional attitudes and emotional perturbations and distinguish between two senses of anger. The second reason draws on people's memory for their past emotional reactions.

In one sense, anger corresponds to a stable attitude to respond aggressively to certain eliciting conditions. Anger, that is, would correspond to a more or less stable emotional feature of one's personality. An irascible person is angry in this sense. In another sense anger is a perturbation: a state whereby an agent has an urge to act aggressively. As a perturbation, anger has relatively short duration, while as an emotional attitude anger is a permanent personality trait. Anger works best as a signaling device when it consists in an emotional perturbation. Agents, in fact, characteristically display certain physical cues such as facial expression, tone of voice and posture when they feel an emotional perturbation. Such cues can indicate an urge to yell out their rage and behave aggressively. But in this sense anger tends to “spend itself” quickly (Frijda 1986, p. 43). After an outburst of anger people tend to calm down and the display of anger fades away even if the conditions that elicited that reaction still remain. So in this sense anger is not an enduring indicator of an agent’s future behaviour. It is not a reliable indicator either since outbursts of anger need not signal irascible agents. Irascible agents have a permanent predisposition to act aggressively, but they don’t display permanently such a commitment. Sugden’s argument requires that agents can reliably recognize what type of individuals they are interacting with; it requires that agents can reliably recognize irascible individuals. But irascible individuals do not permanently display their commitment to aggression.

Irascible people, however, might display the short-term cues fairly often given the right circumstances, and thus they could be reliably identified in fairly small communities. But then they will find themselves shunned and miss opportunities for mutually beneficial interactions with others. As already noted with Elster (1998, p. 72): “They may gain more in each interaction, but interact more rarely. They will not, moreover, be able to learn that their emotional disposition works against them, and hence will have no incentive to control themselves.” Hence, being reliably identified as irascible could not pay off in small communities. The link between the evolutionary advantages of anger as commitment and as signaling device on which Sugden builds his argument for the emergence of resentment seems lost.

The appeal to retrospective evaluation of past injuries may fail to establish the link between anger as a commitment device (which shapes our preferences) and backward-looking resentment for a second reason. If anger as a commitment device doesn't depend in any systematic way on looking back in anger at past injuries, then agents may often prefer *not* to behave aggressively by acting on their remembered anger. For the intensity of the emotion felt by retrospective evaluation of past injuries may often be attenuated, and so past anger may not shape agents' preference towards aggressive behaviour in a durable way. If this is so, then looking back in anger may fail to commit agents to a certain course of action.

Now, there are multiple factors that can influence people's memory for past emotions (Levine et al. 2006, for review). So, memories for past emotional reactions are often inaccurate. Specifically, there is evidence that current appraisals concerning whether or not an individual is responsible or not for negative circumstances predict whether the intensity of remembered anger is over or under-estimated: emotions inconsistent with current appraisals are underestimated (Levine et al. 2001). Furthermore, retrospective evaluation of emotional experience have been found to be explained by a peak-and-end rule according to which people's estimates of past emotional experience can be reliably predicted as the average of the peak emotional intensity and the end emotional intensity of the experience (Fredrickson and Kahneman 1993). Two of the consequences of such rule are that the net (un)pleasantness and how long the experience lasted are not taken into account in our memory for emotional experience. Given two angry affective episodes A and B, adding an extra period of anger to A but not to B will attenuate the intensity of remembered anger in A if the added period ends less angrily. Backward-looking anger, then, may often be attenuated or fade away.

Even assuming that anger is both evolutionary advantageous and a basic emotion, Sugden's conclusion that "backward-looking resentment is the evolutionary price that has to be paid for this advantage" doesn't rest on solid grounds.

2. Hedonism and Norm Compliance

Sugden's argument focuses on negative emotions like anger, fear and resentment. *Even if* his argument fails, it would still be possible that positive emotions are the ultimate motivational source of norm compliance behaviour. Pleasure is the main candidate here, as it has traditionally been linked to motivation.

It has been suggested that data on the neurobiological processes underlying social preference are best understood in hedonic terms. From this perspective, pleasure would be the ultimate motivation for norm compliance. I now defend the claim that the current neurocomputational evidence does not establish a hedonist interpretation.

Let me start by briefly recalling what social preferences are. Theories of social preference model how people rank allocations of material payoff to self and others during strategic interaction (Fehr 2009). According to these theories, individuals are also concerned with the payoff, preferences and beliefs of other individuals. Notice that theories of social preferences are not committed to any specific interpretation of the processes underlying decision-making. In particular they do not make any claim with respect to the hedonic significance of norm compliance behaviour.

Ernst Fehr and Colin Camerer have argued that a hedonic interpretation of theories of social preference provides a good explanatory framework for interpreting the neurobiological data on norm compliance (Fehr and Camerer 2007). They draw on experimental findings from neuroeconomics to support the claim that individuals derive “higher hedonic value” from outcomes associated to the decision to comply with norms of cooperation or fairness (Ibid., p. 420). Fehr and Camerer’s (2007) argument can be reconstructed as follows.

> P1. Norm compliance, in general, and “altruistic, fair and trusting behaviors” in particular, “are consistently associated” with neural activity in the striatum (p. 419).
>
> P2. Activity in the striatum represents anticipated or experienced reward.
>
> C1. Norm compliance is rewarding.
>
> C2. People comply with social norms because it is rewarding.

If we assume that reward is just hedonic value or pleasure, then C2 is a version of motivational hedonism. Motivational hedonism, in its strongest formulation, is the claim that only pleasure (or pain) motivates us. Fehr and Camerer (2007) do not claim that the evidence they review is sufficient to establish C2. Still, they claim that the evidence strongly supports the hypothesis that norm compliance and pro-social behaviour have special reward value. To evaluate Fehr and Camerer’s argument we need answer two questions: First, what does reward amount to here? Second, can the same data considered by Fehr and Camerer be plausibly explained with no appeal to pleasure? After having recalled the types of findings that purportedly support C1, I engage with P2 by considering different computational roles of the striatum. Different meaning of ‘rewards’ are distinguished, and I argue that pleasure does not ground norm compliance.

Fehr and Camerer review evidence from three types of sources. First, they cite an unpublished work by Kosfeld, Fehr, and Weibull where questionnaire data would support the view that "mutual cooperation in social exchanges has special subjective value, beyond the value that is associated with monetary earnings" (p. 420). Second, they survey the findings of a number of neuroimaging experiments where striatum activity has been observed to be significantly correlated with cooperative outcomes. Third, they notice that striatal activity in one experimental condition can be used to predict choice behaviour in a different experimental condition, thereby lending support to C2, that is to the claim that norm compliance occurs because it is rewarding.

Since Fehr and Camerer do not provide details of Kosfeld et al's questionnaire, it is difficult to assess whether, and to what extent, a hedonic component plays a role in the subjects' ratings. There is some evidence that dopamine activity does not most reliably correlate with ratings of the hedonic experience associated with a drug. For example, in spite of significant loss of most dopamine neurons in the basal ganglia, patients with Parkinson's disease have been reported to have normal subjective pleasure ratings for sweet food (Sienkiewicz-Jarosz et al. 2005).

However, the strongest reason provided by Fehr and Camerer in support of hedonic interpretations of theories of social preferences is that other-regarding and norm-compliance behaviour is consistently associated with activation in the striatum. The striatum is part of what is called the "reward circuit." Camerer and Fehr interpret the processes carried out by activity in this area in terms of hedonic processes. But 'reward' and 'hedonic processes' are equivocal. And in light of current evidence the computational role of the striatum is probably complex and may well comprise a number of sub-computational routines, as Chapter 1 suggested. A brief description of the anatomy of the striatum should highlight this last point.

As mentioned in Chapter 1, the striatum is a subcortical part of the brain. It is the main input station of the basal ganglia which are primarily implicated in motor control, learning and decision-making. Because dopamine is the major striatal neuromodulator, the striatum is thought to be one of the main hubs of the reward circuit. However, it is not the only area associated with "reward" processing: the ventral tegmental area, the amygdala, the prefrontal cortex and certain parts of the thalamus are also involved in reward processing. The ventral part of the striatum consists in the caudate nucleus and the putamen. The ventral striatum—or nucleus accumbens—constitutes a third subdivision of the striatum. These three striatal regions are anatomically and functionally distinct. Current evidence suggests that discrete regions of the striatum are differentially involved in the integration of sensorimotor, cognitive and emotional information, and in action selection and initiation (Knutson et al. 2009). But what does it mean that the striatum processes "rewards"?

Here are examples of rewards that seem to be processed by such a circuit are: sweet tastes, cocaine, sex, money, smiling faces, and norm compliance. In a general sense, rewards can be understood as objects or states that make us come back for more. In narrower sense, reward refers to subpersonal informational signals that play specific roles in RL algorithms implemented by certain populations of neurons.

Reward *as* a psychological notion has distinct aspects. The neuroscientist Kent Berridge identifies three dissociable aspects of reward: *liking*, *wanting* and *learning* (e.g. Berridge 2003; Berridge et al. 2009). ‘Liking’ refers to the hedonic experience of a subject. Reward here is a state or outcome that generates a pleasant feeling. ‘Wanting’ (or ‘incentive salience’) refers to a drive towards the pursuit and/or consumption of some, typically salient, state or outcome. It need not be conscious. Reward in this sense is what is desired, often unconsciously, regardless of its hedonic properties. Learning involves the capacity to associate stimuli, and actions to consequences. Reward here consists in states, events and stimuli that guide agents’ learning.

In light of these distinctions, to say that the striatum processes reward can mean at least three different things. Since the relevant sense for Fehr and Camerer’s argument is the hedonic one, we should read P2 above as: Activity in the stratum represents anticipated or experienced pleasure. To assess P2 we should then turn to consider the evidence about the neurobiological underpinnings of hedonic experience.

Berridge and collaborators provide substantial evidence that liking or hedonic experience is generated by opioid, endocannabinoid and GABA-benzodiazepine neurotransmitter systems. Two “hedonic hot-spots” have been found respectively in the nucleus accumbens and in the ventral pallidum which are two regions of the striatum in fact. The first hot-spot comprises 10% of the volume of the nucleus accumbens: a relatively small portion of the striatum. Outside those hot-spots, in the same two regions, opioids do not enhance liking: enhancement of hedonic experience is then anatomically restricted to small portions of the striatum (Berridge and Kringelbach 2008). Hence, the claim that the striatum is a hedonic area need be strongly qualified.

One main reason for interpreting the striatum as a pleasure-centre has been that its major afferents come from dopamine neurons, which have been traditionally considered "pleasure neurotransmitters." However, manipulations of dopamine activity do not appear to have systematic effects on hedonic experience (Berridge and Robinson 1998). This suggests that dopamine activity is neither necessary nor sufficient for generating hedonic experience, but, psychologically, is probably necessary for "wanting," and neurocomputationally—as we have seen—for implementing certain forms of RL algorithms (see Berridge 2007 on "wanting"; Schultz 2007a for distinct computational roles of dopamine).

If hedonic experience is the ultimate motive of norm compliance, pleasure should be the triggering cause of the selection of a certain action. Pleasure should be prior to "wanting" and should determine what we shall do. Evidence needed to confirm or refute this claim might be gained by focusing on the causal relationship between mechanisms of action selection and "liking."

Computational models of the basal ganglia, of which the striatum is the major nucleus, provide one way to approach this issue. Recall that in the framework of reinforcement learning, as Chapter 1 made clear, one of the best models of the basal ganglia mechanism is the Actor-Critic architecture (Houk 2007). This class of models seems to capture important principles of dopaminergically controlled plasticity in the striatum (e.g. Joel et al. 2002). In such models an "Actor" selects the action to be taken given the current input while the "Critic" drives the learning process by assessing how well the outcome of one action tallies with the attainment of a certain goal. Neurobiological data suggests that the ventral striatum is associated with the Critic, while the dorsal striatum is associated with the Actor (Daw, Niv and

Dayan 2005, Sec. 4). Although these computational models are simplistic if compared to the complex anatomy and physiology of the striatum, we can still draw some conclusions about the relationship between action selection and pleasure.

From the characterization of the computational architecture of the striatum, and the localization of hedonic hotspots thereof, it seems that the main computational business of the striatum is not hedonic-processing. Hedonic hotspots *might* be activated *after* the Critic has computed to what extent the outcome of the action taken matches with what was expected. They would register the pleasure of learning rather than driving learning itself. If this is so, then hedonic experience would be the output of decision-making systems over which pleasure has no direct control. The causal interplay between learning and what we want would then make pleasure a contingent result of the appraisal of the outcomes of our actions. Pleasure, therefore, is probably not the ultimate motive of norm compliance: people generally do not comply with norms because it feels good. If the story on offer in this thesis is roughly correct, then people generally comply with norms to minimize sensory- and reward-prediction errors, and sensory and reward-prediction errors should not be identified with emotions agents feel. I now fill in my neurocomputational proposal by focusing on caring.

3. Neurocomputation and Caring

Caring, I have agreed with Frankfurt, is a complex mode of the will. It consists in committing oneself to one's own desire. It corresponds to a relatively stable volitional profile that steadily and persistently orients and guides one's behaviour. I conclude this chapter by proposing that caring might depend on as a fundamental aspect of the RL-system. The mechanism for caring understood in this way might be necessary both to feel emotions and to act on certain social representations so as to comply with norms. Here is my proposal.

The capacity to care about something might depend on particular parameter-values of RL algorithms being in a specific range. Chapter 1 argued that RL neural computations are crucial to agents' capacities to act upon social representations so as to comply with norms. As already pointed out, the proper workings of RL algorithms depend on several parameters (also called *meta-parameters*). The experiment reported in the next chapter attempts to estimate some of these parameters, which may regulate subjects' learning of social norms and decision-making in social contexts. For the moment, let me focus on three important parameters in RL algorithms: the learning rate η, the discount factor γ, and the temperature τ.

The learning rate η controls how quickly old information is updated by experience; for small values of η, learning will be slow, while for large values of η, what has already been learned may be quickly updated. When η is too large, the learning process becomes unstable.

The discount factor γ controls the time scale of reward prediction. More precisely, it determines how far in the future rewards should be taken into account. This is particularly important in case of possible conflicts between immediate and long-term outcomes. The smaller γ, the more the agent will be focused on short-term outcomes only. Too large γ can lead to unreliable predictions of future reward.

The temperature τ controls the randomness of the action choice. Small values of τ promote "explorative" behaviour by which more information is gathered about the mapping of which actions are rewarding. Large values of τ favor "exploitative" behaviour whereby action selection is at making the best use of what has already been learned; as τ tends to infinity all actions have the same probability of being selected.

The setting and adjustment of these parameters are crucial for RL algorithms to successfully carrying out cognitive functions. If agents' capacity to act upon social representations so as to comply with norms is enabled by RL neural computations, then the setting and adjustment of these parameters are crucial to successfully navigate our social environment and comply with norms.

Based on neurobiological data and computational results, Doya (2002) puts forward a hypothesis concerning the role of specific neuromodulatory systems in computing these parameters (see also Schweighofer and Doya 2003). *Neuromodulators* are neurotransmitters that have spatially distributed and temporally extended effects on their receptors. They affect globally and at longer time scale the computations that brains carry out. Doya's hypothesis is that ascending neuromodulatory systems are the media for signaling and adjusting the parameters that regulate the workings of RL systems in the brain with their concerted interaction. Specifically, according to Doya's hypothesis, the acetylcholinergic system controls the learning rate η; the serotonergic controls the discount factor γ; the noradrenergic system controls the temperature τ. Let me now expand a little on some details of Doya's hypothesis.

Acetylcholine seems to modulate synaptic plasticity in the cerebral cortex, stratum, amygdala and hippocampus. Depletion of acetylcholine neurons is associated to memory disorders. So the acetylcholine system, Doya surmises, may modulate "the information coding in the cortex and the hippocampus so that their response properties are not simply determined by the statistics of the sensory input but are also dependent on the importance of the sensory inputs" (Doya 2002, p. 503). The acetylcholine system then might control the learning rate η by affecting the storage and update dynamics of memory at both cellular and circuit levels.

Higher levels of serotonin would determine higher setting of γ which would lead agents to be sensitive to reward predictions longer in the future. Low levels of serotonin instead would lead agents to be insensitive to larger delayed rewards and so to behave impulsively. Serotonin might control the discount factor in RL systems by directly influencing the computations of reward-prediction errors in the basal ganglia, which receive serotonergic input from the dorsal raphe nucleus, or more diffusely by differentially enhancing or inhibiting the activity of parallel RL algorithms which might be implemented in distinct neural populations across the brain.

Noradrenaline is known to be involved in the control of fight-or-flight response and noradrenalinergic neurons are especially active in urgent situations. This is consistent with the idea that noradrenaline regulates the randomness in action selection, which should be sensitive to the urgency of the situation and the stage in learning: higher levels of noradrenaline would determine higher setting of τ which leads to "exploitative behaviour," noradrenaline levels, Doya notices, should decrease when the action value function which determines the agent's decisions has a high variance for a given state.

Now, the complex, concerted neurocomputational activity of these neurotransmitters might, at least partly, determine what an agent cares about. If caring is individuated by whether and when some novel action should be taken instead some old course of action, by how far into the future outcomes of one's action should be taken into account, and by what needs be retained in memory and what can be overwritten, then the concerted neurocomputational activity of these neurotransmitters would underlie the capacity to commit oneself to a certain desire, that is, the capacity to care. Given specific tunings of η, γ, and τ, the progress of one's learning and the structure of the environment, an agent will be committed to a certain desire and tend to act in specific ways, feel certain emotions in determinate circumstances, and think in certain ways rather than others. The mechanism for caring understood in this way might be necessary both to feel emotions and to act on certain social representations so as to comply with norms and behave smoothly and co-adaptively with other agents in the social environment.

Conclusion

This chapter has filled in the neurocomputational mechanism of norm compliance on offer in this thesis with another detail. It characterized the capacity to care in terms of the concerted setting and adjustment of specific RL-parameters. Such parameter-dynamics might be underlain by particular neuromodulatory systems. It argued that emotion is probably not the ultimate motive of norm compliance behaviour: caring might be the source of both feeling certain emotions and complying with norms. The focused on the notion of reward in RL-computation, and argued that hedonic interpretations of rewards are unjustified.

The next and last chapter of this thesis will elaborate further on the relationship between reward and norm compliance behaviour. It will address the questions of whether and how the nature of the rewards received by people after they make decisions in a social situation affects their propensity to learn a social norm. The chapter will present experimental and computational results relevant to those questions, and provide me with the opportunity to put some of the ideas explored in this thesis at work.

CHAPTER 7.

Social Rewards and Learning Social Norms. An Experimental Study

(Joint work with Aistis Stankevicius and Peggy Seriès)

The goals of this chapter are twofold. First, the chapter will illustrate how some of the concepts and modelling tools used in previous chapters can be put to work. Second, it will address two questions concerning the relationship between rewards and the learning of social norms by reporting and discussing the results of an experimental project, to which I have contributed. In so doing, it will clarify, by means of experimental data, the impact of distinct types of rewards on people's motivation to comply with norms, which was explored both in the introduction and in the previous chapter. Specifically, the questions addressed here are:

(i) Does the type of the reward outcomes obtained by people after they make decisions in social situations affect the way they learn a social norm?

(ii) When people are learning a social norm, how do *social* reward outcomes, as opposed to *non-social* reward outcomes, affect their decision-making processes?

The chapter has four sections. The first presents and motivates alternative hypotheses concerning questions (i) and (ii). The second reports the results of a model-based experiment I have collaborated to. The third section discusses these results. The fourth concludes.

1. Social and Non-Social Rewards

Consider question (i). One hypothesis is that the type of reward outcomes (or feedback cues) *per se* does not have significant impact on learning and social decision-making because both social and non-social reward outcomes would be processed in the same way by the same neural circuit. Support for this hypothesis comes from two classes of findings in neuroeconomics. First, there seems to be substantial overlap between the neural circuits active in tasks where behaviour is guided by social norms, and neural activations observed in reinforcement learning tasks (for reviews see Fehr and Camerer 2007; Lee 2008). Second, both social and non-social reward outcomes—including money, food, juice, facial expressions and verbal feedback—engage overlapping neural substrates during reinforcement learning computations (e.g. Delgado et al. 2000; Berns et al. 2001; O'Doherty et al. 2002; Walter et al. 2005; Behrens et al. 2008; Izuma et al. 2008; Spreckelmeyer et al. 2009; Lin et al 2011). These types of neurobiological findings suggest, therefore, that the type of reward outcomes *per se* should not make any significant difference on the way people learn social norms.

An alternative hypothesis is that different types of reward outcomes (or feedback cues) have different impact on learning and decision-making. If widely different, both social and non-social environmental cues have large impact on people's social behaviour, then different types of reward outcomes may also differently affect the way people change their behaviour to adapt to novel social situations. A substantial amount of evidence from experimental economics and social psychology indicates that in fact social behaviour is sensitive to very subtle situational cues. For example, people are more likely to litter in a particular environment when it is heavily littered than when the same environment is clean (Cialdini et al. 1990; Cialdini 2003). Showing experimental participants a picture of a library and instructing them to go to the library after the experiment can lead them to whisper during the experiment (Aarts and Dijksterhuis 2003). Finding a dime in the coin return slot of a public telephone makes it twenty-two times more likely that one will help a woman who has dropped a folder full of papers (Isen and Levin 1972). More relevant here, contributions to public goods tend to increase when people make decisions while they are "watched" by a pair of eyes drawn on an honesty box (Bateson et al 2006; see also Haley and Fessler 2005; Rigdon et al 2009).

Findings such as these demonstrate, therefore, a strong influence of apparently insignificant cues in the environment on social behaviour (see e.g. Bargh and Williams 2006; Doris 2002 for a critical review). This suggests that different types of reward outcomes (or feedback cues) observed multiple times in a given social situation may differently affect the way a social norm is learned in that situation.

Consider question (ii) now: when people are learning a social norm, how could *social* reward outcomes, as opposed to *non-social* reward outcomes, affect their decision-making processes? Some of the studies reviewed above indicate that social cues can have substantial impact on decision-making. Emotional expressions can systematically bias learning processes and decision-making (Averbeck and Duchaine 2009; Evans et al. 2011). Evidence indicates that social cues can increase pro-social behaviour: images of a pair of eyes can significantly increase cooperative behaviour not only in a laboratory condition, but also in real-world contexts (Bateson et al. 2006; Ernest-Jones et al. 2011). With respect to learning performance, some studies on a feedback-guided item-category association task show that learning is more effective when the feedback provided to participants consists of facial expressions of emotion (happy or angry faces) instead of non-social cues such as red and green lights, and that this effect is supported by brain regions such as the amygdala, distinct from the dopamine-based reward circuit (Hurlemann et al. 2010; Mihov et al. 2010). Results from a recent study also indicate that feedback information provided by some social cues is processed by different neural circuits than non-social, cognitive feedback (Evans et al. 2011).

These findings, therefore, underwrite the hypotheses that social cues significantly affect cooperative behaviour and social reward outcomes, in comparison to non-social reward outcomes, often enhance learning performance. These effects would be mediated by brain regions besides dopamine-based reward circuits.

In light of this body of evidence, relevant to address the two questions above, three hypotheses were tested in the present study. First, different types of reward outcomes have different effects on learning and social-decision making. Specifically, social reward outcomes have a different impact on learning and social decision-making than non-social reward outcomes. Second, when compared to non-social cognitive feedback, social reward outcomes in the form of facial expressions lead participants to display more pro-social behaviour. Third, when compared to participants who are provided with non-social cognitive feedback, participants receiving feedback in the form of facial expressions learn a social norm more effectively.

The study presented in what follows tested these hypotheses by using an associative learning task, which we called "the tipping game." The task allowed us to examine the effects of social reward outcomes, as opposed to non-social reward outcomes, on learning and decision-making.

2. The Tipping-Game

2.1 Methods

Participants

Forty participants (17 females), between the ages of 19 and 37 (Mean = 26.22; Standard Deviation = 4.31), performed a decision-making task. The majority of the participants were students in the University of Edinburgh recruited through an internal university mailing list. All participants signed informed consent and were compensated with £6/hour for taking part in the experiment. The study was approved by the University of Edinburgh, School of Informatics Ethics Committee.

Task

Participants were initially given five short questionnaires to fill in: the “Empathy Quotient” (EQ) questionnaire (Baron-Cohen and Wheelwright 2004), one version of the “Reading the Mind in the Eyes” test (Baron-Cohen et al. 1997), the “Self Report Altruism” questionnaire (Rushton et al. 1981), the “Sensitivity to Punishment and Sensitivity to Reward Questionnaire” (SPSRQ) (Torrubia et al. 2001), and the “Behavioural Inhibition/Approach” (BIS/BAS) questionnaire (Carver and White 1994). These questionnaires measured the levels of empathy, mentalizing, altruism, and punishment and reward sensitivity of the participants.

The aims of collecting questionnaire data about participant’s personality traits were twofold. Firstly, we were interested in understanding whether performance in the tipping game could be explained solely by some relatively stable personal trait, rather than by the feedback provided. Secondly, information about personality traits could be used to better characterize qualitatively the nature of the behavioural results observed.

Once they completed the questionnaires, participants took part in a decision-making task. In the task, participants were instructed to pretend that they were visiting a foreign country far-away, and that they repeatedly were going to dine at restaurants. They were endowed with fictional monetary units (*mu*), with which they had to pay for restaurant bills and for any tip they decided to leave. Their goal was to learn how much they were expected to tip at the end of a meal in that country without spending too much money. The goal, put differently, was to learn a social norm of tipping so as to display adaptive behaviour in the social situation they were facing (the exact instructions given to participants are reported in Appendix A, at the end of the chapter).

To motivate participants to pursue this goal, they were informed at the beginning of the task that the best performance would be rewarded with £20 and that this performance would be measured in function of both how well the social norm was learned and how much fictional money (*mu*) was saved.

The task represented tipping situations as sequential interactions, where a server chooses the service quality, and then the diner chooses the tip. After each decision, a reward outcome is revealed and can be used by the diner to learn how to adapt to the social situation she or he is facing (Figure 1).

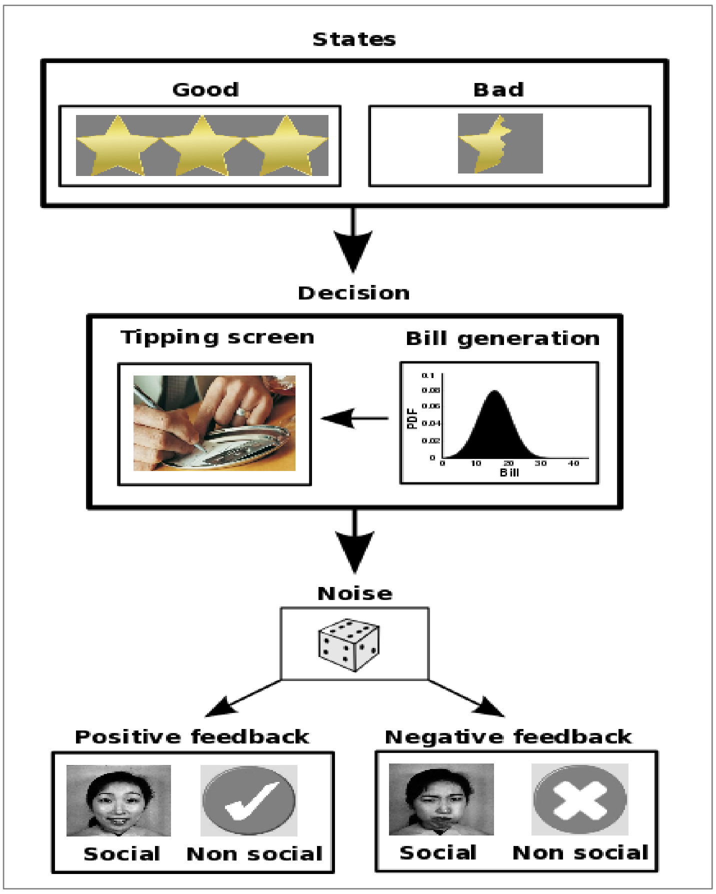


Figure 1. Sequence of events during one trial. The state of the environment is initially revealed: it corresponds either to good or to bad service received at a restaurant. A decision screen follows, which informs the participant about how much money *mu* is available and how much the bill is. The participant is then asked to make a decision about how much he or she wants to tip. The last screen provides feedback. The feedback depends stochastically on state-action pairs and the underlying social norm of tipping. In the social condition, the feedback consists of either happy or angry faces, while in the non-social condition it consisted of a tick or a cross mark.

The task consisted of three blocks each of which comprised forty trials. At the beginning of each block, participants were endowed with 1,100 *mu*. For each trial, the service quality could be either good or bad. In each trial across the three blocks, the chance of getting good service was 0.5. After the service quality was revealed, participants were informed about how much *mu* they had left and the amount of the bill they had to pay. Bills were drawn from a distribution with mean 18 and standard deviation 5, truncated to [3, 45]. Participants were then asked to make decisions about how much they wanted to tip. Any sum equal or greater than zero could be tipped. The bill and the amount tipped were subtracted from this endowment after each trial so that participants could keep track of the *mu* they were spending. After participants made a decision, either positive or negative feedback was shown. This ended a trial.

Participants were informed at the beginning of the task that there could be some manipulations across blocks. In fact, two types of manipulations took place. The first type of manipulation consisted in changing the underlying social norm of tipping. In the first block the social norm of tipping was 23% of the bill. In the second block the social norm was 50%. In the third block it was 23% again.

The second type of manipulation across blocks was the variation in the reliability of the reward outcomes: the feedback provided had different levels of noise. Reward outcomes, in fact, depended stochastically on the underlying social norm of tipping and on the pair service quality-amount tipped (Table 1).

	Action (tip < norm / tip ≥ norm**)**		
	Block 1 (Norm: 23%)	**Block 2** (Norm: 50%)	**Block 3** (Norm 23%)
Good State	20/80	20/80	35/65
Bad State	30/70	30/70	40/60

Table 1. Mapping from state-action pairs and underlying norm to outcomes.
Numbers in the cells refer to the chance (in percentage) of obtaining a *positive* reward-outcome, which was a function of the state observed, the underlying norm of tipping and the action taken. In each cell, the first number refers to the chance of obtaining a positive reward outcome when the action taken was less than the social norm of tipping; the second number refers to the chance of obtaining a positive reward outcome when the action taken was greater or equal to the underlying social norm.

In the first and second blocks, if the service quality was good and the amount tipped by the participant was equal or greater than the social norm, then there was an 80% chance to receive a positive reward outcome and a 20% chance to receive a negative reward outcome. If the service quality was good, but the amount tipped was less than the social norm, then there was a 20% chance to receive a positive reward outcome and an 80% chance to receive a negative reward outcome.

If the service quality was bad and the amount tipped was equal or greater than the social norm, then there was a 70% chance to receive a positive reward outcome and a 30% chance to receive a negative reward outcome. If the service quality was bad and the amount tipped was less than the norm, then there was a 30% chance to receive a positive reward outcome and a 70% chance to receive a negative reward outcome.

In the third block, if the service quality was good and the amount tipped was equal or greater than the social norm, then there was a 65% chance to receive a positive reward outcome and a 35% chance to receive a negative reward outcome. If the service quality was good, but the amount tipped was less than the norm, then there was a 35% chance to get a positive reward outcome and a 65% chance to get a negative reward outcome. If the service quality was bad and the amount tipped was equal or greater than the norm, then there was a 60% chance to receive a positive reward outcome and a 40% chance to receive a negative reward outcome. If the service quality was bad but the amount tipped was less than the norm, then there was a 40% chance to receive a positive reward outcome and a 60% chance to receive a negative reward outcome.

The between-subjects independent variable in the task was the type of the reward outcomes provided to the participants after they made a decision about how much to tip. In one condition (Social Condition), twenty participants (7 females), between the ages of 20 and 33 (Mean = 25.7; Standard Deviation = 3.82), received feedback in the form of a happy or an angry face. In a second condition (Non-social Condition), twenty participants (10 females), between the ages of 19 and 37 (Mean = 26.75; Standard Deviation = 4.78), received non-social feedback in the form of a tick or an *X* mark after their decisions.

In the Social Condition, two types of facial expressions were used: one had a happy expression and the other had an angry expression. Two different identities for the facial expressions were also used; they were alternated pseudo-randomly across the three blocks. Four pictures were used in total: two happy facial expressions and two angry facial expressions (Figure 2). The pictures were selected from the Japanese Female Facial Expression (JAFFE) database (Lyons et al 1998).

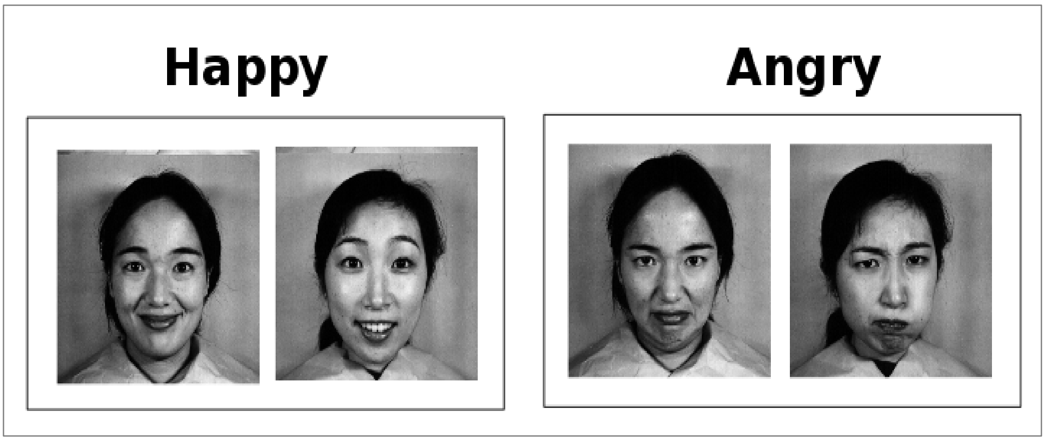
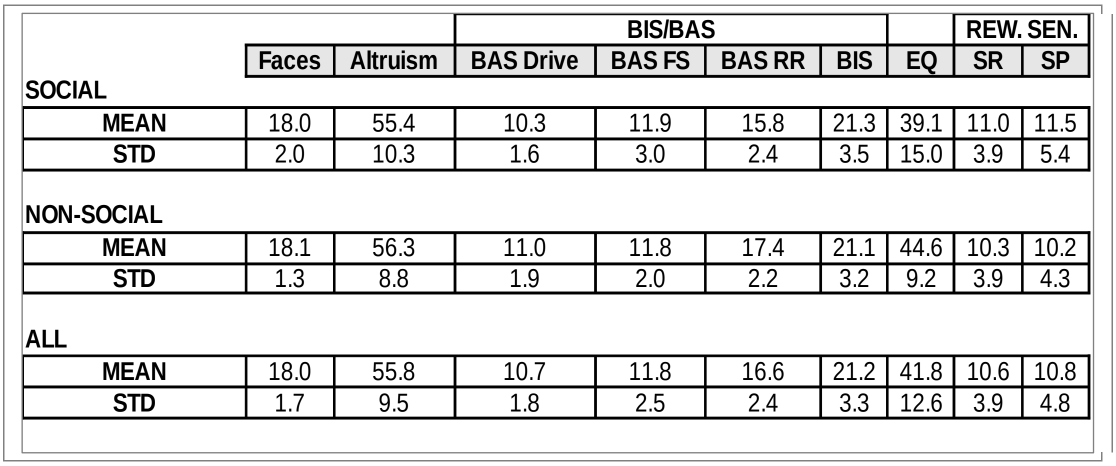


Figure 2. Feedback stimuli in the social condition. In the social condition positive feedback consisted of happy faces, while negative feedback consisted of angry faces. Feedback stimuli were selected from the Japanese Female Facial Expression (JAFFE) database (Lyons et al. 1998).

All participants displayed normal capacity for recognition of facial expressions. On average, they also presented normal levels of altruism, empathy and attitudes towards reward and punishment as measured by the personality questionnaires they answered (Table 2).



			BIS/BAS					REW. SEN.	
	Faces	Altruism	BAS Drive	BAS FS	BAS RR	BIS	EQ	SR	SP
SOCIAL									
MEAN	18.0	55.4	10.3	11.9	15.8	21.3	39.1	11.0	11.5
STD	2.0	10.3	1.6	3.0	2.4	3.5	15.0	3.9	5.4
NON-SOCIAL									
MEAN	18.1	56.3	11.0	11.8	17.4	21.1	44.6	10.3	10.2
STD	1.3	8.8	1.9	2.0	2.2	3.2	9.2	3.9	4.3
ALL									
MEAN	18.0	55.8	10.7	11.8	16.6	21.2	41.8	10.6	10.8
STD	1.7	9.5	1.8	2.5	2.4	3.3	12.6	3.9	4.8

Table 2. Average scores per group. Table entries indicate average scores for “Reading the Mind in the Eyes” (Faces), “Self Report Altruism”, “Behavioural Inhibition/Approach” (BIS/BAS) questionnaires, “Empathy Quotient” (EQ), “Sensitivity to Punishment and Sensitivity to Reward Questionnaire” (SPSRQ). All scores are in “normal” ranges.

In the Non-social Condition, two types of symbols were used: a tick (also known as a check mark or check) and an *X* mark. The tick is a symbol generally used to indicate that the action taken is good or correct. The *X* mark, instead, generally indicates that the action taken is bad or incorrect. Although there are cross-cultural differences in the way the tick mark and the *X* mark are understood, all participants in the present experiment stated in a debriefing questionnaire administered after the task that they recognised the symbols as the notation respectively for “good” (or “correct”) and for “bad” (or “incorrect”) (the debriefing questionnaire is found in Appendix B, at the end of this chapter).

2.2 Results

From the debriefing questionnaires, it was found that one participant reported that he had not understood the task. This participant's results were then excluded from data analysis. To examine our hypotheses, paired two tailed t-tests were used. The t-tests were run on data concerning the decisions of 39 participants (20 participants for the Social Condition; 19 participants for the Non-social condition) averaged for each trial.

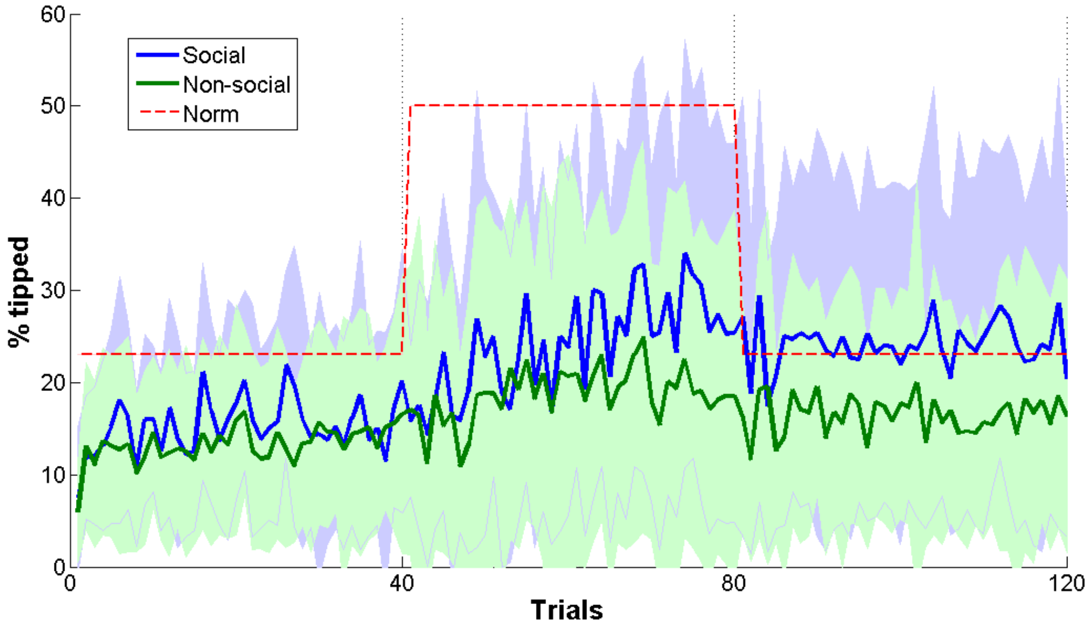


Figure 3. Comparison between participants in the social and non-social condition with respect to the average amount tipped over trials in the three blocks in the task. The blue and green lines correspond to the average amounts (as percentage of the bill) tipped per trial by participants receiving respectively social and non-social feedback. The dotted red line corresponds to the underlying social norm of tipping (as percentage of the bill). Blue and green shades refer to standard deviations from the group-average amounts tipped.

As shown in Figures 3-4, social reward outcomes were found to have a different impact on learning and social decision-making than non-social reward outcomes, thus confirming our first hypothesis. Specifically, it was found that: first, participants in the social condition tipped significantly more than participants in the non-social condition (across all blocks, mean difference = 4.17%, p < 0.001; in Block 1, mean difference = 0.99%, p = 0.043 ; in Block 2, mean difference = 4.94%, p < 0.001; in Block 3, mean difference = 8.19%, p = 0.019).

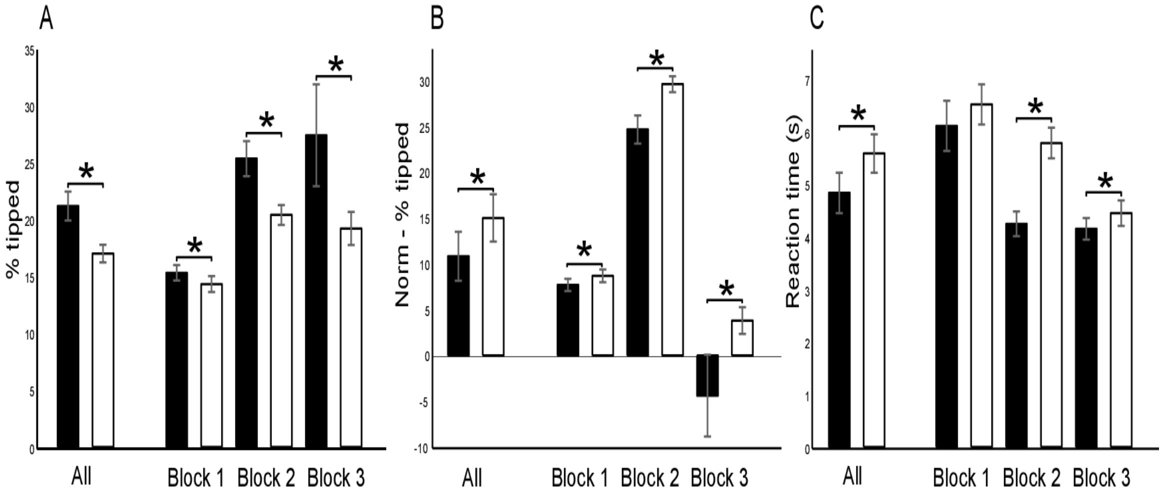


Figure 4. A: Amount tipped in the social (black bars) and non-social (white bars) condition. The leftmost histogram pair shows the average amount tipped by participants in the social and non-social groups with respect to the whole task. The other three pairs of histograms show results with respect to single blocks. *B*: Absolute difference between the social norm and the amount tipped in the social and non-social condition. The leftmost histogram pair shows results over the whole task. The other three pairs of histograms display results for single blocks. *C*: Time (in seconds) taken to make decisions. The leftmost histogram pair shows results over the whole task. The other three display results for single blocks.
Asterisk indicates that the difference between the two groups is statistically significant (p-value < 0.05).

Second, the absolute difference between the social norm and the amount tipped by participants in the social condition was significantly lower than the absolute difference between the social norm and the amount tipped by participants in the non-social condition (across all block, mean difference = - 4.17%, p < 0.001; in Block 1, mean difference = - 0.99%, p = 0.043; in Block 2, mean difference = - 4.94%, p < 0.001; in Block 3, mean difference = - 8.19%, p = 0.019).

Third, participants in the social condition made significantly quicker decisions than participants in the non-social condition (across all blocks, mean difference = - 749 ms, $p < 0.001$; in Block 1, mean difference = - 409 ms, $p = 0.176$; in Block 2, mean difference = -1538 ms, $p < 0.001$; in Block 3, mean difference = - 301 ms, $p = 0.047$).

These three differences were observed across all blocks. The first finding confirms the second hypothesis: social reward outcomes in the form of facial expressions led participants to display higher degree of pro-social behaviour. The second and third findings confirm the third hypothesis: learning is facilitated by social feedback.

A large inter-individual variability of performance was observed at the task, with some subjects learning to adapt their decisions to conform to the norm much better than other subjects, who seemed to behave independently of feedback throughout the experiment (Figure 5).

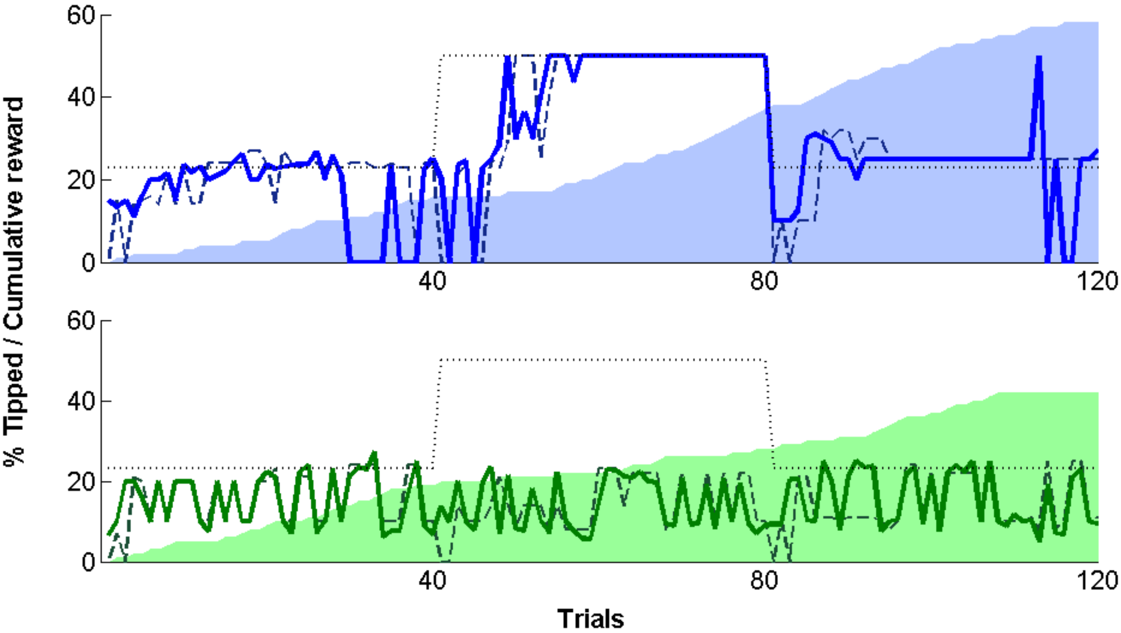


Figure 5. Two examples of subjects' learning performance: subject 10 (social group) at the top, subject 35 (non-social group) at the bottom. Blue or green solid curves represent the percentages tipped, coloured areas in the background represent the cumulative rewards received and dotted black curve is the underlying norm of tipping. Subject 10 is an example of a good performer: the subject displayed significant changes in behaviour towards the social norm across blocks. Subject 35 is an example of a bad performer, as the behaviour the subject displayed across blocks does not significantly change towards the social norm. Coloured dashed curves represent percentages tipped by the corresponding model. The actions with maximum *Q*-values are plotted to show how much actions with the maximum likelihood differ from the ones chosen by the subject.

With respect to our third hypothesis, we observed that significantly more participants displayed learning in the social condition than participants in the non-social condition. More specifically, to identify the number of learners in the task, we assumed that if the mean amount tipped by a participant was significant different from one block to the next and moved towards the underlying social norm, then that participant displayed learning. According to this criterion, we found that 12/20 displayed learning in the social group, vs. 7/19 in the non-social group (Figure 6).

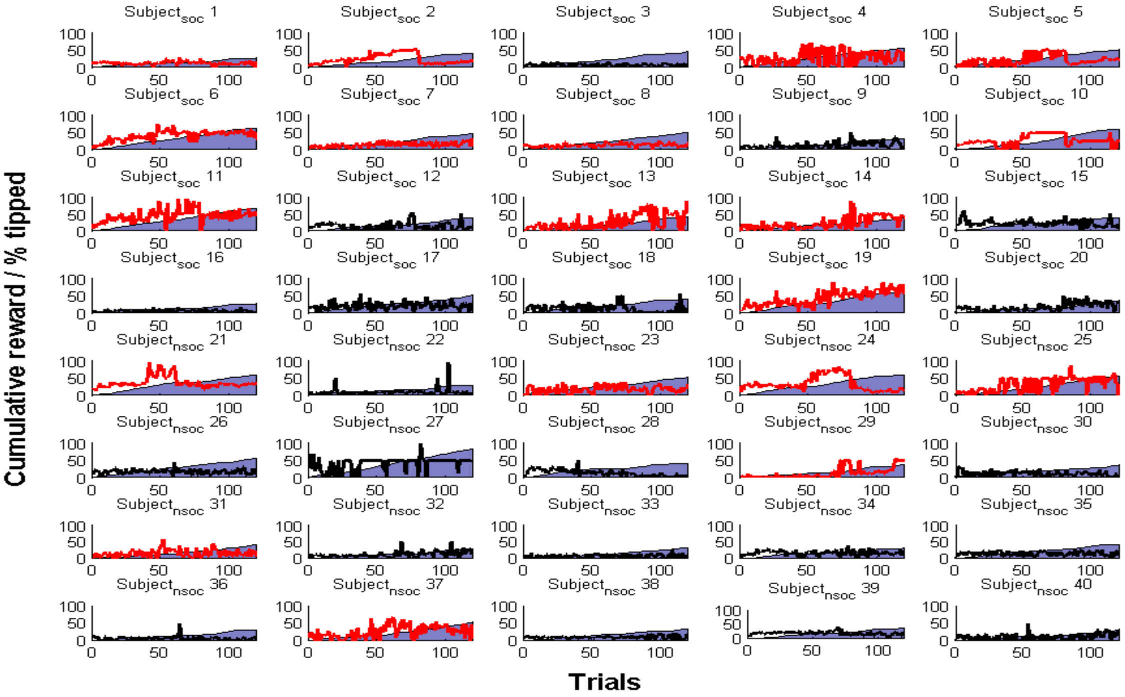


Figure 6. Individual data plots for tipping percentages, cumulative rewards and learning criteria satisfaction. Red (good learners) or black (bad learners) coloured curves represent the percentages tipped, and blue curves show the cumulative rewards received.

Two further points should be noted about the behavioural results that we observed. First, although the absolute difference between the social norm and the amount tipped by participants in the social condition was significantly lower than that for participants in the non-social condition, in the first two blocks, participants in both groups failed to learn the underlying social norms: they always tipped much less than the norm. One hypothesis is that they had a strong prior bias towards a specific action different from the social norms in our task. Second, standard deviations from average percentages tipped in both groups were high, especially in the third block (see Figure 1). In the first block, standard deviations were respectively 2.99 *mu* and 3.03 *mu* for the social and non-social group respectively. In the second block, they were: 6.9 and 3.79, while in the third they were 20 and 6.34. Finally, considering the whole experiment, the standard deviation from the average percentage tipped by the social group was 5.6402 *mu*, vs. 3.3948 *mu* for the non-social group. So, in general, we observed high variability in the behavioural data; in particular, participants in the social group took actions, which were more spread out over a larger range of values in comparison to the actions taken by participants in the non-social group.

2.3 Model

To further describe quantitatively the nature of the effects that we observed, we explored whether the behaviour of participants could be modeled with a type of Rescorla-Wagner reinforcement learning algorithm (Rescorla and Wagner 1972). The model algorithm could make decisions in our task with the goal of maximizing its total reward. It could do this by learning action values *Q* for state-action pairs, and selecting, at each trial, actions in function of their estimated *Q*-values.

The possible states were two, corresponding to "good" or "bad" service quality. The action space comprised 101 actions, corresponding to tip percentages from 0% increasing in steps of 1% to 100%. For each of the two states, the action taken by the model was assigned a value, which was a function of both the reward outcome obtained for taking that action, the economic cost incurred, and the *Q*-value of that state-action pair stored in memory. This is expressed by the *Q*-update equation:

$$[1] \qquad Q(\text{state, action})_{\text{new}} = Q(\text{state, action})_{\text{old}} + \alpha(\text{reward}- Q(\text{state, action})_{\text{old}})$$

where α is the learning rate ($0 \leq \alpha \leq 1$), which determines the learning step-size, that is, how fast learning takes place. The smaller α, the least the existing knowledge is modified. Conversely, as α tends to 1, what has already been learned can be quickly overwritten.

The action selection mechanism was governed by a softmax function. At any given trial, the model chose action *a* from among the possible actions with probability:

$$[2] \qquad Prob\left(a_t \mid state_t\right) = \frac{e^{\tau Q\left(a_t, state_t\right)}}{\sum_{j=0}^{N} e^{\tau Q\left(a_j, state_t\right)}}$$

where *N* is the number of actions the agent can take. τ is a positive parameter called inverse temperature. As τ tends to ∞, the action with highest *Q*-value has a much higher probability of being selected than the others. As τ tends to 0, all actions become equally probable.

Given the questions and hypotheses that motivated our study, we focused on the reward signal in our model. The reward consisted of the weighted average of two components: an economic component and a reward outcome component. Formally:

$$[3] \qquad Reward = \frac{r_{out} w_{out} + r_{econ} w_{econ}}{w_{\max}}$$

where r_{econ} is an economic factor and is equal to the Tip/Bill ratio, which could take any value in the interval [0, 1]. The economic weight w_{econ} ($-w_{max} \leq w_{econ} \leq w_{max}$) is a parameter that determined to what extent spending money was valued in the tipping game. If w_{econ} was $-\ w_{max}$, then spending money was valued very negatively, thereby characterizing a type of agent with a “stingy” attitude; if w_{econ} was w_{max}, then spending money was valued very positively, thereby characterizing a type of agent with a "generous" attitude. r_{out} is a reward outcome factor, which was associated to the two possible outcomes in the task: either positive feedback or negative feedback. The reward magnitude of both happy facial expressions and the tick mark was assumed to be 1. The reward magnitude of both angry facial expressions and the *X* mark was assumed to be – 1. The outcome weight w_{out} ($-w_{max} \leq w_{out} \leq w_{max}$) is a parameter that determined to what extent positive feedback was valued in the tipping game. If w_{out} was – w_{max}, then positive feedback was valued very negatively, and negative feedback was valued very positively. Agents with a negative w_{out} could be characterized as "punishment-seeking" types. If w_{out} was w_{max}, then positive feedback was valued very positively, and negative feedback was valued very negatively. Agents with a positive w_{out} could be characterized as "reward-seeking" types. Four types of agents could be distinguished in function of the values of the two reward parameters, that is, in function of their attitudes towards economic costs and reward outcomes.

2.4 Modelling Results

To estimate parameters values, for each block in the task we fitted the model to participants' data using maximum likelihood estimation. For each participant, we rounded each of his or her action to the corresponding entry in the *Q*-table of the action space. Each entry in the *Q*-table represented the value of selecting a particular action *a* for a given state *s* in our task. The actions that our model could take were "clamped" to the actions taken by each of our participants. We then fitted the model by using maximum likelihood estimation. By searching the parameter space, we found the set of parameters that maximized the likelihood of each participant's observed sequence of actions.

Our model could describe the behavioural data reasonably well, for both group and individual performance (Figure 5 for two examples). For the social condition, the mean difference between percentages tipped by participants and those tipped by the fitted models was 4.2% (standard error 0.16%). For the non-social condition, the mean difference was 4.4% (standard error 0.17%).

The parameter values thus obtained confirmed that participants in the social group displayed learning and decision-making profiles different from the non-social group. Specifically, across blocks average parameter values governing learning and decision-making for participants in the social condition differed from parameter values for participants in the non-social condition (Table 3 A). Such variation indicates that participants' learning and decision-making were affected by the nature of the feedback outcomes received as well as by the changes in the underlying norm and reliability of the feedback that took place across blocks.

Considering the modelling results for the whole task, instead of per individual blocks, participants in the social condition displayed on average higher learning rate α and economic weight w_{econ} than participants in the non-social condition. On the contrary, they displayed smaller inverse temperature τ and outcome weight w_{out} (Table 3 B). These results indicate that on average, in comparison to participants in the non-social condition, participants in the social condition: learned more quickly, explored more actions, sought positive feedback less, and cared less about spending extra *mu* for tips.

		α	τ	w_{out}	w_{econ}
Block 1	**Social**	0.56	4.80	2.15	5.125
	Non Social	0.46	8.50	1.725	5.30
Block 2	**Social**	0.44	6.50	1.225	4.025
	Non Social	0.53	5.60	2.275	6.375
Block 3	**Social**	0.60	4.40	2.375	4.05
	Non Social	0.48	5.20	2.80	6.15

Table 3. A). Average parameter values per block for participants in the social and non-social condition.

		α	τ	w_{out}	w_{econ}
Tipping Game	**Social**	0.42	3.60	2.575	5.75
	Non Social	0.31	5.80	3.225	4.00

Table 3. B). Average parameter values per experiment for participants in the social and non-social condition

The most significant differences between the two groups concerned the learning rate α and the inverse temperature parameter τ. Except for the second block in the task, the learning rate was significantly greater for participants in the social group than for participants in the non-social group. Except for the second block, the average values of the temperature parameter τ for participants in the social group were significantly smaller than those for participants in the non-social group. These two findings were confirmed by modelling results for the whole task.

From results concerning mean values of the weights w_{out} and w_{econ}, it was found that four participants (three in the social condition) were “stingy reward-seekers” (i.e. w_{econ} was negative, while w_{out} was positive), two (one per group) were generous punishment-seekers” (i.e. w_{econ} was positive, while w_{out} was negative), the rest of the participants in the task were of the “generous reward-seeker” type (i.e. both parameter weights w_{out} and w_{econ} were greater than zero). So, attitudes towards economic cost and reward outcomes were not abnormal. This was independently confirmed by questionnaires results, given the “normal range” of our participants' personality scores (Table 2 above).

Finally, the hypothesis that the behavioural effects observed in the tipping game could be explained solely by some stable personality trait was ruled out. In fact, no pattern of significant correlations was found between questionnaire scores one the one hand, and behavioural and modelling results on the other.

3. Discussion

Our study asked whether and how the type of the reward outcomes obtained by people after they make decisions in social situations affects the way they learn a social norm. We addressed these questions by determining whether the influence of facial expressions on participants' decisions in an associative learning task, called the “tipping game”, was significantly different from the influence of non-social feedback in the form of conventional marks. We found that participants receiving feedback in the form of happy or angry facial expressions behaved significantly different from participants receiving feedback in the form of tick or cross marks. This effect was observed across all the blocks in our task, and, specifically, had impact on: how much participants were willing to give as a tip, how well they learned the underlying social norm and how fast they made decisions.

Interestingly, results about reaction time together with self-reported information about the strategy used to make decisions indicate that participants' decision-making processes were distinctively affected by the type of reward outcomes received (Appendix C at the end of this chapter). Unlike participants in the non-social condition, nearly all participants in the social condition stated that they relied on the feedback provided, either positive or negative, without attempting to work out the right amount they were expected to tip. Thus, in comparison to participants in the non-social condition, their learning and decision-making relied more on quick, unconscious, and apparently more effective processes.

In blocks one and two, on average, no participant learned the value of the underlying social norm. The most significant differences between groups were observed in blocks two and three. One plausible explanation for this finding is that, when they started the task, participants had an initial bias in favor of a specific amount that one should leave as a tip in restaurants. Results from the debriefing questionnaire (Appendix C) indicate that in fact most of the participants had specific expectations about tipping in restaurants, namely: they generally expected that tips should be in the range of 15-20% of the bill. As confirmed by the average amount tipped in the first block, participants may have initially relied more heavily on such prior expectation. Systematic exposure to feedback stimuli may have then gradually overcome the effect of this initial bias, and led participants to acquire new beliefs about how much one should tip in the situations they faced in our task.

Taken together, our findings are *prima facie* inconsistent with the hypothesis that the type of reward outcomes *per se* does not have significant impact on learning and social decision-making. It should be pointed out, however, that this conclusion holds only if we assume that the magnitudes of the reward outcomes in the two conditions of our experiment were perfectly matched. Based on behavioural results alone, it might be granted that angry faces and cross marks were aversive and that happy faces and tick marks were appetitive. However, one might hypothesize that their magnitudes were different, so that participants found more rewarding viewing a happy face than a tick mark (or more punishing viewing an angry face than a cross mark). Thus, in comparison to tick and cross marks, viewing angry and happy facial expressions could have had more impact on the computations driving learning and decision-making because of their differential magnitudes, and not because of their social or non-social nature—in other words, they could have more impact because r_{out} would be greater in the case of facial expressions. This would be consistent with the hypothesis that because all types of reward outcomes are processed through a common circuitry different types of reward outcomes *per se* do not make significant difference in learning and decision-making.

Whether different types of reward outcomes are perfectly matched for magnitude cannot be determined easily using behavioural results (see Evans et al. 2011). It is important to notice, however, that although shared neural circuits might be involved in the computation of both social and non-social reward outcomes—as suggested, for example, by Lin et al. (2011) and Jones et al (2011)—the full network involved in processing both types of reward outcomes is probably not identical. Hence, when considered in the context of neuroimaging studies investigating the impact of social stimuli on reward-based decision processes (Evans et al. 2011; Lin et al. 2011; Walter et al. 2005), and if we take into account results about reaction times in our task, our findings provide increasing support to the hypothesis that social reward outcomes bias learning and decision-making differently from non-social outcomes.

Our modelling results provided one possible way to quantitatively characterize this bias whose effects were observed in the behaviour of our participants. The parameter values that we estimated suggest that obtaining social, instead of non-social, reward outcomes may have greater impact on (1) the rate to which newly acquired information overrides old knowledge and (2) the tendency to explore more of the action space available. According to our modelling results, in fact, the behavioural differences observed between groups in our task were better accounted for by differences in their rate of learning and action selection strategy than by differences in the attitudes that participants could have towards different types of reward outcomes. This conclusion was independently underwritten by our questionnaires results, where no significant difference was found between the two groups with respect to their level of empathy, altruism, and sensitivity to rewards and punishments.

Feedback in the form of facial expressions could lead people, who are learning a social norm in a new environment, to adapt more effectively to the social situation they are facing. Angry facial expressions, in particular, might drive such learning by affecting the decision-making strategy underlying social behaviour.

Angry facial expressions might signal social disapproval of a failure to comply with a certain norm. Such a failure might be due to a lack of knowledge of the social environment. Thus, learners of a social norm might feel anxious and uncomfortable in observing an angry reaction, which might draw their attention to their ignorance of the structure of the social situation they are facing (on the role of punishment on the emergence of norms of cooperation see e.g. Fehr and Gächter 2002). Interestingly, the desire to avoid social disapproval is in fact one of the main factors that may motivate people to tip in restaurants (Azar 2007b; Conlin et al 2003).

If discomfort is to be avoided and knowledge of how one ought to act is to be acquired in that situation, then people should, at least initially, sample extensively the action space by trying many different actions until an accurate representation of the environment is gained. Even after people are confident that they have come to possess accurate knowledge of the environment, it could still be effective to trying new actions occasionally. Using this type of action selection strategy, people would make sure that nothing has changed in the structure of the environment. This is especially important in social situations, also involving social norms of tipping, where new social norms can appear, existing social norms change and old ones disappear relatively quickly across places and over time (see Azar 2004a, 2004b on the evolution of tipping).

Accordingly, the tendency of our participants in the social condition to display a less "greedy" action selection strategy could be explained if social negative feedback was especially effective in drawing their attention to the need to gain a better representation of the structure of the environment. Awareness of their need to have accurate knowledge of the situation along with a desire to steer clear from social disapproval could have stimulated a willingness to explore a bigger portion of the action space available. By exploring more actions, participants could improve estimates of non-greedy action values and, at the same time, display on average more generous behaviour. Ultimately, exploration along with a relatively higher learning rate could lead participants in the social condition to better adapt to the situation they were facing.

Two limitations of our study should be noted before we conclude. The first concerns the distinction between social and emotional cues. The stimuli that we used in the social condition of our task did not help us to determine whether the behavioural effects we observed depended on social rather than on only the emotional dimension of facial expressions. Facial expressions are in fact means to convey both social and emotional information. Besides communicating information about other agents, facial expressions can often elicit emotional reactions in the observers. In order to identify the role of emotional cues alone, in contrast to facial expressions, on participants' learning and social decision-making, a third condition for our task may employ emotional, non-social reward outcomes.

Second, one reason why our subjects did not generally perform well in the tipping game might be that its reward structure made the learning task especially hard. The level of noise in the mapping between state-action pairs and reward outcomes was high across the three blocks, making the feedback provided not very reliable. Moreover, the reliability of the feedback was independent from the distance between amount tipped and underlying social norm, so that tips well above the social norm could still receive negative feedback outcomes. In order to improve learning performance, the reward structure of the task may be modified in two ways. On the one hand, the level of noise in the mapping between state-action pairs and reward outcomes may be diminished across all blocks. On the other hand, the reliability of the feedback provided may be made dependant on the distance between amount tipped and underlying social norm, so as to strengthen the reliability of feedback outcomes for tips well above or well below the social norm.

Conclusion

Our study confirmed the hypothesis that different types of reward outcomes differentially affect the way people learn a social norm and make-decisions. Results from our tipping game demonstrated that social reward outcomes in the form of facial expressions, if compared to non-social reward outcomes in the form of conventional feedback marks, can lead people to learn more effectively a social norm of tipping. Specifically, social reward outcomes in the form of facial expressions can lead people to make relatively quicker and more pro-social decisions, and ultimately to adapt more easily to novel social situations. In order to explore quantitatively our participants' behaviour, we used a version of the Rescorla-Wagner algorithm to model performance in the tipping game. Modelling results suggest that the different pattern of performance between participants in the social and non-social condition could be better explained by a drive, displayed by participants in the non-social condition, to acquire knowledge by trying more novel actions.

Appendix A: Instructions provided to participants of the Tipping Game

Imagine that you are a stranger just arrived in a foreign country. You believe that when people go to a restaurant they normally leave a tip at the end of their meal. You also believe that how much people tip in a restaurant depends on the quality of the service they receive in that restaurant. But you may be wrong. You are going to have a number of meals at restaurants in that new country. Imagine that this is the type of situation you are about to face in this experiment.

The experiment consists of three parts. There are 40 trials in each part. Each trial corresponds to a meal that you have in a restaurant while you are in that country. Each part of the experiment corresponds to a new visit to that country. So each time you go to that country you have 40 meals at restaurants. You believe that some things might have changed since your last visit. But you are not sure.

Imagine that every time you arrive in that country you have *mu* 1,100 in your pocket – *mu* is the local currency. You need to use this money to pay your bills and for any tip you wish to leave at the end of your meals.

In each trial in the experiment, you will initially be revealed the quality of the service you receive in the restaurant. Here is an example:

[service quality screen]

You will then be reminded how much you have left in your pocket, and you will be presented with your bill. You will be asked how much you wish to leave as a tip in that situation on top of your bill. You can tip any sum greater than or equal to zero by pressing the appropriate keys on the keyboard. When you have decided press ENTER to confirm your decision.

Here is an example:

[decision screen]

After your decision you will receive some feedback.

[face or symbol]

This is an example of feedback you may receive.

In each part of the experiment, do your best to adapt your behaviour to the new situation without spending too much money.

At the end of the experiment you will receive a score based on how well you have adapted in that type of social situation in that country *and* on the amount of money left in your pocket.

You will receive an extra prize in cash depending on your score in the task.

Remember that you are making a non-negligible contribution to science… and that you have the chance to win extra money. Thanks in advance for your participation and attention.

RECAP

- You have to imagine that you are a stranger just arrived in a foreign country, and that you are going out for dinners in that country.

- The service quality in that restaurant is revealed.

- Imagine you have eaten your dinner and you pay the bill.

- You decide how much you want to leave as a tip.

- Some feedback is displayed.

Appendix B: Debriefing Questionnaire

1. Did you find the first, second or the third part harder, or were they both the same?

1st part harder   2nd part harder   3rd part harder   about the same

2. Did you notice any change across the three parts of the experiment?

Yes   No

3. If you answered yes in the question above, then please explain what changes you noticed between:

1st and 2nd part:____

2nd and 3rd part:____

1st and 3rd part:____

4. What do you think was the social norm of tipping (please give a percentage of the bill, e.g. 32% of the bill):

1st part:___

2nd part:___

3rd part:___

5. What do these signs mean?

✓ _____
X _____

6. Did you use the feedback provided to make your decisions?

YES   NO

7. Which of the following descriptions best describes the strategy you used to make your decisions? (tick statement that you most agree with)

a. I relied on the feedback and I tried to work out the right percentage.

b. I relied on the feedback provided without working out the right percentage.

c. I relied on the positive feedback mostly without spending too much time thinking.

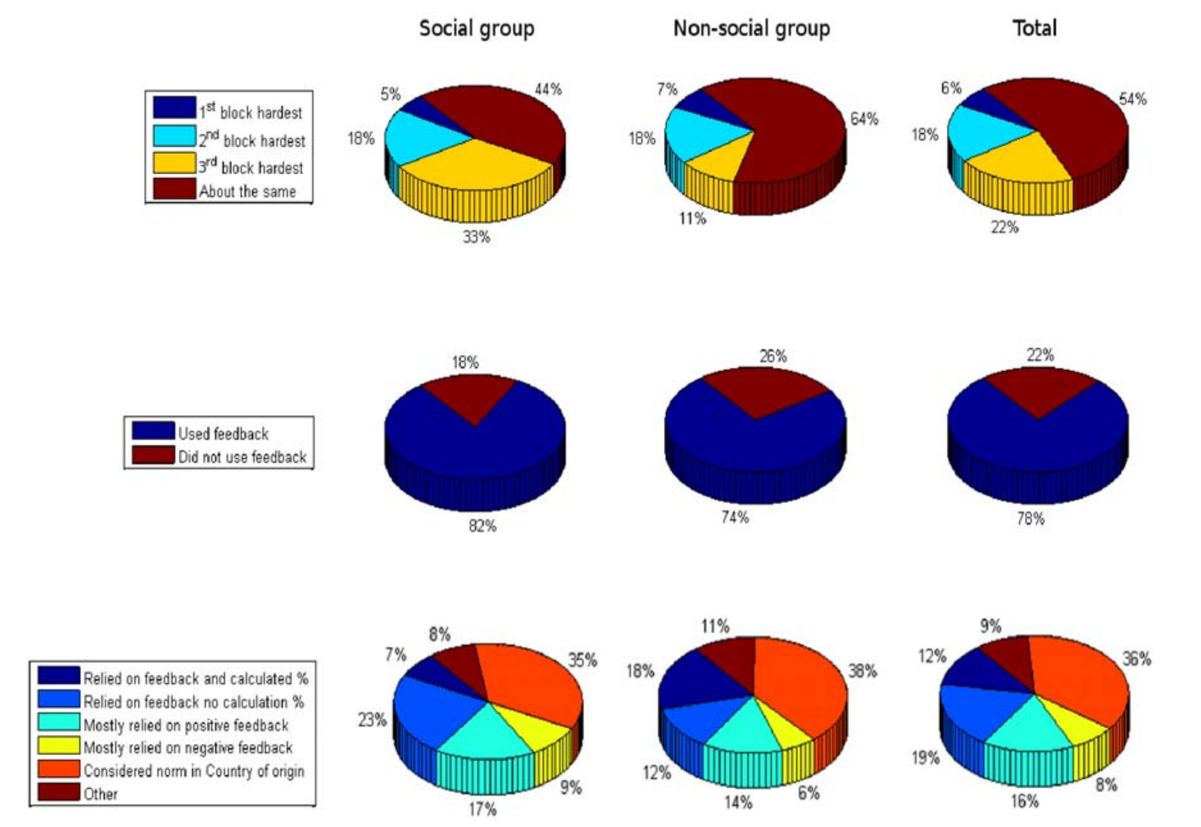
d. I relied on the negative feedback mostly without spending too much time thinking.

e. I considered the norm of tipping in my country. (Please indicate your country of origin:__)

f. Don’t know.

g. Other. (Please give a short description): _____

Appendix C: Debriefing Questionnaires Statistics



CONCLUSIONS

This neurocomputational investigation into social norm compliance began with a triplet of questions and answers. The triplet—recall—goes like this:

Q: How can we make progress in our understanding of social norms and norm compliance?

A: Adopting a neurocomputational framework is one effective way to make progress in our understanding of social norms and norm compliance.

Q: What could the neurocomputational mechanism of social norm compliance be?

A: The mechanism of norm compliance probably consists of Bayesian-Reinforcement Learning algorithms implemented by activity in certain neural populations.

Q: What could information about this mechanism tell us about social norms and social norm compliance behaviour?

A: Information about this mechanism tells us that:

a_1 Social norms are uncertainty-minimizing devices.

a_2 Social norm compliance is one trick we have devised to interact co-adaptively and smoothly in our social environment.

This journey now concludes by considering each of the Q-As in light of the claims articulated and defended by the previous chapters.

Progress with respect to research questions such as "What are social norms? And why do people comply with them?" is due to empirical discoveries, mathematical advances, but also to development of new theoretical frameworks. Establishing a novel framework to studies of human normativity is in fact a significant contribution in itself.

The first main claim defended by this thesis is that social norms and norm compliance can be effectively understood within a neurocomputational framework, whereby the workings of the mechanism of social and moral behaviour can be identified and described. Analytical tools and concepts from fields such as statistical decision theory, machine learning, computer science and reinforcement learning have been increasingly used to make sense of data about the neural bases of social norm compliance. The marriage between theoretical approaches and experimental research in social neuroscience has helped to unify results from such disciplines as philosophy, economics, anthropology, psychology and artificial intelligence, to articulate more sophisticated theories of social behaviour and to address more complex empirical problems concerning norm compliance in a precise and reliable way. In the last section of Chapter 1, these reasons were given in support of the claim that our understanding of social norms and norm compliance can make effective progress if we examine social and moral behaviour within a neurocomputational framework. Chapters 2 and 3 argued, more specifically, that adopting a neurocomputational perspective is fruitful to understanding whether (Chapter 2) and how (Chapter 3) explanations of social norm compliance should appeal to representations; Chapter 6 argued that a neurocomputational perspective can help us to identify the motivational structure of norm compliance. Finally, the experiment described in Chapter 7 attempted to showcase some of the fruits that a neurocomputational exploration of norm compliance can yield. My hope is that future research will vindicate the claims I put forward in those chapters.

Bayesian decision theory and Reinforcement Learning have proved successful in uncovering important features of the mechanisms of perception and action. Drawing upon such successes, the second main claim advanced in the thesis is that the building blocks of the mechanism of social norm compliance probably consist of Bayesian and Reinforcement Learning algorithms running on certain neural circuits. The suggestion is that social/moral behaviour piggybacks on neural computations that enable agents to process incoming sensory input so as to form probabilistic beliefs about the states of the world causing that input, and to choose actions so as to maximize the value of their future reward outcomes in the social world. Thus, social norms could be grounded in features of human nature, which are more fundamental than either the beliefs and preferences of individuals or the idiosyncratic characteristics of the culture in which they live.

Chapter 1 laid down the beginnings of such a neurocomputational model of norm compliance and pointed to possible neural circuits for perception and action in the social/moral domain. The concerted activity of these circuits would be geared towards minimizing uncertainty over interactions with other agents in the social environment. Chapters 4, 5 and 6 articulated particular aspects of the model put forward in Chapter 1. Putative Bayesian, explanatory ingredients of the mechanism of normative judgement were considered in Chapter 4. Chapter 5 examined the relationship between language and moral cognition, and suggested that the peculiar "norm-hungriness" of humans is dependent on the capacity for florid-control, which might be enabled by neural computations executed by basal ganglia-prefrontal cortex activations. In focusing on the relationship between emotion and the motivational structure of norm compliance, Chapter 6 argued that the capacity to care, which is essential to motivate norm compliance, is enabled by certain interactions between specific neuromodulators. The dynamics of these neuromodulators—the chapter claimed—might correspond to specific settings of the parameters that control Reinforcement Learning algorithms.

A full neurocomputational account of social norm compliance—it should be clear—is far from being simple. Here, I point to two important challenges. First, if the neural system carries out Bayesian and Reinforcement Learning algorithms so as to enable norm compliance, then such algorithms must run quickly and efficiently. Rapid adaptation to changes in real-world social circumstances often requires that the learning of new pieces of social knowledge and that the decision of whether one ought to comply with a certain social norm should be "thoughtless" and effortless. Yet, Bayesian computations seem to be too resource demanding, especially in the social domain, where hidden states of the environment are extremely high-dimensional and continuous. Moreover, Reinforcement Learning algorithms are often too slow when confronted with real-world situations where the number of possible states and actions that an agent can take is huge. This means that there are two key challenges for a descriptively adequate neurocomputational model of social norm compliance. One challenge is to identify appropriate forms of *approximate* Bayesian inference; the other challenge is to explore more *sophisticated* learning algorithms, which could operate quickly upon suitable representations of the environment. Approximate Bayesian inferences and sophisticated learning algorithms might enable us to deal with the complexity of the social world, while making feasible demands on our resource-bounded brains. These types of algorithmic models should be explored systematically in the context of social navigation.

The second challenge for an adequate neurocomputational model of norm compliance is to identify algorithms with richer dynamical interactions between perceptual, motivational-valuation and control systems. As Gershman and Daw (Forthcoming) put it: “Perception, action and utility are ensnared in a tangled skein.” Although the model I put forward in this thesis may suggest that perception and action underlying norm compliance are supported by separate signals, with a clean separation between inference-driven perception and reward-based action selection, it is likely that (social) perception is in fact modulated through and through by reward-information. Accordingly, motivational-valuation and perceptual systems may not consist of separate, dedicated neurocomputational mechanisms. As research on the neurocomputational foundations of social norms proceeds, it is plausible that the building blocks of the mechanism of social norm compliance will include algorithms beyond “pure” Bayesian and Reinforcement Learning ones.

If the mechanistic model I have proposed is roughly on the right track, there are two properties that appear to be essential to social norms and social norm compliance. Social norms would be uncertainty-minimizing devices and social norm compliance would be one of the tricks we can employ to interact co-adaptively and smoothly in our social environment. These two properties are uncovered by the mathematical concepts from statistical decision theory, which I have used to investigate the neurocomputational foundations of norm compliance. Accordingly, the notions of “uncertainty” and “management of social uncertainty” would be crucial to describe and make sense of social norm compliance. If this is so, then the concepts we use to account for why people comply with norms should be informed by the fact that norms are intimately related to social uncertainty. Chapters 1 and 4 explained in which sense uncertainty is bounded up with social norms and moral judgement. Chapters 5 and 6 were partly concerned on how co-adaptive and smooth interaction is facilitated by norm compliance.

One way to summarize the thrust of the argument developed over these chapters is with Mary Douglas's words:

> "Institutional structures [can be seen as] forms of informational complexity. Past experience is encapsulated in an institution's rules, so that it acts as a guide to what to expect from the future. The more fully the institutions encode expectations, the more they put uncertainty under control, with the further effect that behavior tends to conform to the institutional matrix […]. They start with rules of thumb, and norms; eventually, they can end by storing all the useful information" (Douglas 1986, p. 48).

From a neurocomputational perspective, the idea is that by minimizing uncertainty over their social interactions, agents' cognitive systems become models of the social environment in which the agents are embedded. To perceive our social world would then be to successfully predict our own sensory states brought about by social states. A normative system can be understood as one device for communicating, sharing and acting upon information concerning states in our social landscape. The more a social norm is entrenched in a society, the less computing is needed in order to take the right action. To comply with norms would then be one means to make social predictions come true at little computational cost, so that we can “thoughtlessly” occupy high-valued, low-uncertainty states in our social landscape.

Human life and thought exhibit a range of normative features. One cluster of such features can be brought under the head of social/moral normativity. Humans are first and foremost social creatures who are deeply concerned about what is right or wrong and tend to care for the people with whom they interact. These are among the most theoretically intriguing and practically important characteristics of human life. Understanding social norms and social norm compliance in terms that allow us to see them as aspects of the natural world is a challenging as well as fascinating project, whose significance cannot be overestimated. My neurocomputational journey into norm compliance—I hope—constitutes a step forward towards the realization of that project.

Bibliography

Aarts, H., and Dijksterhuis, A. (2003). “The silence of the library: Environment, situational norms, and social behavior.” *Journal of Personality and Social Psychology*, 84, 18–28.

Abbott, L.F. (2008). “Theoretical neuroscience rising.” *Neuron*, 60, 489–495.

Adolphs, R. (2010). “Conceptual challenges and directions for social neuroscience.” *Neuron*, 65, 752-767.

Anderson, E. (2000). “Beyond *Homo Economicus*: New Developments in Theories of Social Norms.” *Philosophy and Public Affairs*, 29, 170-200.

Anderson, S.W., Bechara, A., Damasio, H., Tranel, D., and Damasio, A.R. (1999). “Impairment of social and moral behavior related to early damage in human prefrontal cortex.” *Nature*, 2, 1032-1037.

Andreoni, J., Harbaugh, W., and Vesterlund, L. (2003). “The Carrot or the Stick: Rewards, Punishments, and Cooperation.” *American Economic Review*, 93, 893-902.

Ashby, F.G., and Maddox, W.T. (2011). “Human category learning 2.0.” *Annals of the New York Academy of Sciences*, 1224, 147–161.

Averbeck, B.B., and Duchaine, B. (2009). “Integration of social and utilitarian factors in decision making.” *Emotion*, 9, 599–608.

Azar, O.H. (2007a). “The social norm of tipping: A review.” *Journal of Applied Social Psychology*, 37, 380-402.

Azar, O.H. (2007b). “Why pay extra? Tipping and the importance of social norms and feelings in economic theory.” *Journal of Socio-Economics*, 36, 250–265.

Azar, O.H. (2004a). “What sustains social norms and how they evolve? The case of tipping.” *Journal of Economic Behavior and Organization*, 54, 49-64.

Azar, O.H. (2004b). “The history of tipping--from sixteenth-century England to United States in the 1910s.” *Journal of Socio-Economics*, 33, 745-764.

Bargh, J.A., and Williams, E.L. (2006). “The automaticity of social life.” *Current directions in psychological science*, 15, 1-4.

Baron-Cohen, S. (2000). “Theory of Mind and Autism: A Fifteen Year Review.” In S. Baron-Cohen, H. Tager-Flusberg, and D.J. Cohen (Eds.), *Understanding Other Minds: Perspectives from Developmental Cognitive Neuroscience*. Oxford: Oxford University Press, 1–20.

Baron-Cohen, S., and Wheelwright, S. (2004). "The Empathy Quotient (EQ). An investigation of adults with Asperger Syndrome or High Functioning Autism, and normal sex differences." *Journal of Autism and Developmental Disorders*, 34, 163-175.

Baron-Cohen, S., Wheelwright, S., and Jolliffe, T. (1997). "Is there a "language of the eyes"? Evidence from normal adults and adults with autism or Asperger syndrome." *Visual Cognition*, 4, 311-331.

Bateson, M., Nettle, D., and Roberts, G. (2006). "Cues of being watched enhance cooperation in a real-world setting." *Biology Letters*, 2, 412-414.

Bayer, H.M., and Glimcher P.W. (2005). "Midbrain dopamine neurons encode a quantitative reward prediction error signal." *Neuron*, 47, 129–141.

Beer, R. D. (2008). *The dynamics of brain-body-environment systems: A status report*. In P. Calvo and A. Gomila (Eds.), *Handbook of Cognitive Science: An Embodied Approach*. San Diego: Elsevier, 99-120.

Behrens, T.E., Hunt, L.T., and Rushworth, M.F. (2009). "The computation of social behavior." *Science*, 324, 1160–1164.

Behrens, T.E., Hunt, L.T., Woolrich, M.W., and Rushworth, M.F. (2008). "Associative Learning of Social Value." *Nature*, 456, 245-249.

Bennett, M.R., and Hacker, P.M.S. (2003). *Philosophical Foundations of Neuroscience*. Oxford: Blackwell.

Berkes, P., Orbán, G., Lengyel, M., and Fiser, J. (2011). "Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment." *Science*, 331, 83-87.

Bermúdez, J.L. (2005). *Philosophy of psychology: A contemporary introduction*. New York: Routledge.

Bermúdez, J.L. (2003). *Thinking without words*. New York: Oxford University Press.

Bernhard, H., Fehr, E. and Fischbacher, U. (2006). "Group Affiliation and Altruistic Norm Enforcement." *American Economic Review*, 96, 217-221.

Berns, G.S., McClure, S.M., Pagnoni, G., and Montague, P.R. (2001). "Predictability modulates human brain response to reward." *Journal of Neuroscience*, 21, 2793–2798.

Berridge, K.C. (2007). "The debate over dopamine's role in reward: the case for incentive salience." *Psychopharmacology*, 191, 391-431.

Berridge, K.C. (2003). "Pleasures of the brain." *Brain and Cognition*, 52, 106-128.

Berridge, K.C., Robinson, T.E., and Aldridge, J.W. (2009). "Dissecting components of reward: 'liking', 'wanting', and learning." *Current Opinion in Pharmacology*, 9, 65-73.

Berridge, K.C., and Kringelbach, M.L. (2008). "Affective neuroscience of pleasure: reward in humans and animals." *Psychopharmacology*, 199, 457-480.

Berridge, K.C., and Robinson, T.E. (1998). "What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience?" *Brain Research Reviews*, 28, 309-369.

Bicchieri, C. (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms.* New York: Cambridge University Press.

Bicchieri, C., and Chavez, A. (2010). "Behaving as Expected: Public Information and Fairness Norms." *Journal of Behavioral Decision Making*, 23, 161–178.

Bicchieri, C., and Xiao, E. (2009). "Do the right thing: But only if others do so." *Journal of Behavioral Decision Making*, 22, 191-208.

Binmore, K. (1994). *Game Theory and the Social Contract, Vol. I. Playing Fair.* Cambridge, MA: MIT Press.

Blackford, J.U., Buckholtz, J.W., Avery, S.N., and Zald, D.H. (2010). "A unique role for the amygdala in novelty detection." *Neuroimage*, 50, 1188–1193.

Blair, R.J.R. (2007). "The amygdala and ventromedial prefrontal cortex in morality and psychopathy." *Trends in Cognitive Sciences*, 11, 387–392.

Blair, R.J.R. (2003). "Neurobiological basis of psychopathy." *British Journal of Psychiatry*, 182, 5–7.

Blair, R.J.R. (1997). "Moral reasoning in the child with psychopathic tendencies." *Personality and Individual Differences*, 22, 731–739.

Blair, R.J.R. (1996). "Brief report: Morality in the autistic child." *Journal of Autism and Developmental Disorders*, 26, 571-579.

Blair, R.J.R. (1995). "A cognitive developmental approach to morality: investigating the psychopath." *Cognition*, 57, 1-29.

Blair, R.J.R., Mitchell, D., and Blair, K. (2005). *The psychopath: Emotion and the Brain*. Malden, MA: Blackwell.

Blair, R.J.R., Colledge, E., Murray, L., and Mitchell, D. (2001). "A selective impairment in the processing of sad and fearful expressions in children with psychopathic tendencies." *Journal of Abnormal Child Psychology*, 29, 491–498.

Blair, R.J.R., Jones, L., Clark, F., and Smith, M. (1997). "The Psychopathic Individual: A Lack of Responsiveness to Distress Cues?" *Psychophysiology*, 34, 192–198.

Blair, R.J.R., Sellars, C., Strickland, I., Clark, F., Williams, A.O., Smith, M., and Jones, L. (1995a). "Emotion attributions in the psychopath." *Personality and Individual Differences*, 19, 431–437.

Blair, R.J.R., Jones. L., Clark. F., and Smith. M. (1995b). "Is the psychopath 'morally insane'?" *Personality and Individual Differences*, 19, 741-752.

Bovet, D., and Washburn, D.A. (2003). "Rhesus macaques (*Macaca mulatta*) categorize unknown conspecifics according to their dominance relations." *Journal of Comparative Psychology*, 117, 400–405.

Camerer, C. (2003). *Behavioral Game Theory: Experiments on Strategic Interaction,* Princeton, NJ: Princeton University Press.

Camerer, C., Teck Hua Ho, T., and Chong, K. (2004). "A Cognitive Hierarchy Model of One-Shot Games." *Quarterly Journal of Economics*, 119, 861-898.

Camras. L. A. (1992). "Expressive development and basic emotions." *Cognition and Emotion*, 6, 269-283.

Carruthers, P. (2002). "The cognitive functions of language." *Behavioral and Brain Sciences*, 25, 657–726.

Carver, C.S., and White, T.L. (1994). "Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales." *Journal of Personality and Social Psychology*, 67, 319-333.

Casebeer, W.D., and Churchland, P.S. (2003). "The neural mechanisms of moral cognition: A multiple-aspect approach to moral judgment and decision-making." *Biology and Philosophy*, 18, 169–194.

Charness, G., and Gneezy, U. (2008). "What's in a name? Anonymity and social distance in dictator and ultimatum games." *Journal of Economic Behavior and Organization*, 68, 29–35.

Chen, Y., and Xin Li, S. (2009). "Group Identity and Social Preferences." *American Economic Review*, 99, 431-457.

Chomsky, N. (1980). "Rules and Representations." *Behavioral and Brain Sciences*, 3, 1-61.

Churchland, P.M. (2000). “Rules, Know-How, and the Future of Moral Cognition.” In R. Campbell and B. Hunter (Eds.), *Moral Epistemology Naturalized: Canadian Journal of Philosophy*, Supplementary Volume XXVI.

Churchland, P.M. (1998). “Towards a Cognitive Neurobiology of the Moral Virtues.” *Topoi*, 17, 83–96.

Churchland, P.M. (1996). “The Neural Representation of the Social World.” In L. May, M. Friedman and A. Clark (Eds.), *Mind and Morals: Essays on Cognitive Science and Ethics*. Cambridge, MA: MIT Press.

Churchland, P.M. (1995). *The Engine of Reason, the Seat of the Soul: A Philosophical Journey into the Brain*. Cambridge, MA: MIT Press.

Churchland, P.M. (1981). “Eliminative Materialism and the Propositional Attitudes.” *Journal of Philosophy*, 78, 67-90.

Churchland, P.S. (2011). *Braintrust: What Neuroscience Tells us about Morality*. Princeton, NJ: Princeton University Press.

Churchland, P.S. (1993). “The Co-Evolutionary Research Ideology.” In A. Goldman (Ed.), *Readings in Philosophy and Cognitive Science*. Cambridge, MA: MIT Press, pp. 745–767.

Churchland, P.S. (1986). *Neurophilosophy: Toward a Unified Science of the Mind-Brain.* Cambridge, MA: MIT Press.

Churchland, P.S., and Sejnowski, T. (1992). *The Computational Brain.* Cambridge, MA: MIT Press.

Cialdini, R. (2003). “Crafting normative messages to protect the environment.” *Current Directions in Psychological Science*, 12, 105–109.

Cialdini, R., and Goldstein, N.J. (2004). “Social influence: Compliance and conformity.” *Annual Review of Psychology*, 55, 591–622.

Cialdini, R., Reno, R. R., and Kallgren, C. A. (1990). “A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places.” *Journal of Personality and Social Psychology*, 58, 1015-1026.

Clapin, H. (Ed.) (2002). *Philosophy of Mental Representation*. Oxford: Oxford University Press.

Clark, A. (2006a). “Material symbols.” *Philosophical Psychology*, 19, 291–307.

Clark, A. (2006b). “Language embodiment and the cognitive niche.” *Trends in Cognitive Science*, 10, 370–374.

Clark, A. (2002a). “The Roots of Norm-Hungriness. Response to John Haugeland.” In H. Clapin (Ed.) *Philosophy of Mental Representation*, Oxford: Oxford University Press, 37-43 and discussion, 44-61.

Clark, A. (2002b). “Skills, spills, and the nature of mindful action,” *Phenomenology and the Cognitive Sciences*, 1, 385-387.

Clark, A. (2000a). “Word and action: reconciling rules and know-how in moral cognition.” In R. Campbell and B. Hunter (Eds.), *Moral Epistemology Naturalized: Canadian Journal of Philosophy*, Supplementary Volume XXVI.

Clark, A. (2000b). “Making moral space. A reply to Churchland.” In R. Campbell and B. Hunter (Eds.), *Moral Epistemology Naturalized: Canadian Journal of Philosophy*, Supplementary Volume XXVI.

Clark, A. 1998: Magic words: How language augments human computation. In P. Carruthers and J. Boucher (eds.), *Language and thought: Interdisciplinary themes*. Cambridge, UK: Cambridge University Press.

Clark, A. (1997). *Being there: Putting brain, body and world together again*. Cambridge, MA: MIT Press.

Clark, A. (1996). “Connectionism, Moral Cognition and Collaborative Problem Solving.” In L. May, M. Friedman and A. Clark (Eds.), *Mind and Morals: Essays on Cognitive Science and Ethics*. Cambridge, MA: MIT Press.

Clark, A. (1989). *Microcognition: Philosophy*, *Cognitive Science*, *and Parallel Distributed Processing*. London: MIT Press.

Clark, A., and Thornton, C. (1997). “Trading spaces: Computation, representation, and limits of uninformed learning.” *Behavioral and Brain Sciences*, 20, 57–90.

Clark, A., and Toribio J. (1994). “Doing without representing?” *Synthese*, 101, 401–431.

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

Cohen, J.D., Braver, T.S., and Brown, J.W. (2002). “Computational perspectives on dopamine function in prefrontal cortex.” *Current Opinion in Neurobiology*, 12, 223–229.

Conlin, M., Lynn, M., and O’Donoghue, T. (2003), “The norm of restaurant tipping.” *Journal of Economic Behavior & Organization*, 5, 297-321.

Cottrell, G.W., and Metcalfe, J. (1991). “Empath: Face, gender and emotion recognition using holons.” In R.P. Lippman, J. Moody, and D.S. Touretzky

(Eds.), *Advances in Neural Information Processing Systems 3*, San Mateo: Morgan Kaufmann, 564–571.

Dana, J., Weber, R.A., and Kuang, X. (2007). “Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness.” *Economic Theory*, 33, 67–80.

Darley, J., and Batson, C.D. (1973). “From Jerusalem to Jericho: A study of situational and dispositional variables in helping behaviour.” *Journal of Personality and Social Psychology*, 27, 100-108.

Dasser, V. (1988). “A social concept in Java monkeys.” *Animal Behaviour*, 36, 225-230.

Daw, N.D., Niv, Y., and Dayan, P. (2005). “Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control.” *Nature Neuroscience*, 8, 1704–1711.

Daw, N.D., Kakade, S., and Dayan, P. (2002). “Opponent interactions between serotonin and dopamine.” *Neural Networks*, 15, 603–616.

Dayan, P. (2009). “Goal-directed control and its antipodes.” *Neural Networks*, 22, 213–219.

Dayan, P., and Niv, Y. (2008). “Reinforcement learning: The Good, The Bad and The Ugly.” *Current Opinion in Neurobiology*, 18, 185–196.

Dayan, P., and Abbott, L. (2001). *Theoretical Neuroscience*. Cambridge, MA: MIT Press.

Dayan, P., Hinton, G.E., Neal, R.M., and Zemel, R.S. (1995). “The Helmholtz machine.” *Neural Computation*, 7, 889–904.

deCharms, R.C., and Zador A. (2000). “Neural representation and the cortical code.” *Annual Review Neuroscience*, 23, 613–647.

Delgado, M.R., Nystrom, L.E., Fissell, C., Noll, D.C., and Fiez, J.A. (2000). “Tracking the haemodynamic response to reward and punishment in the striatum.” *Journal of Neurophysiology*, 84, 3072–3077.

Dennett, D. (1991). *Consciousness Explained*. Boston: Little Brown.

Dennett, D. (1982/83). “Styles of Mental Representation.” *Proceedings of the Aristotelian Society, New Series*, LXXXIII, 213-226.

de Sousa, R. (2010). “Emotion.” *The Stanford Encyclopedia of Philosophy (Spring 2010 Edition)*, Edward N. Zalta (Ed.), URL = <http://plato.stanford.edu/archives/spr2010/entries/emotion/>.

de Waal, F.B.M. (1996). *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Cambridge, MA: Harvard University Press.

de Waal F.B.M., and Tyack P.L. (Eds.) (2003). *Animal Social Complexity: Intelligence, Culture and Individualized Societies.* Cambridge, MA: Harvard University Press.

Dickinson, A. (1985). "Actions and habits: The development of behavioural autonomy." *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, 308, 67-78.

Dickinson, A., and Balleine, B.W. (2002). "The role of learning in the operation of motivational systems." In C.R. Gallistel (Ed.), *Learning, motivation and emotion Vol. 3*. New York: John Wiley & Sons, 497-533.

Dimitrov, M., Phipps, M., Zahn, T.P., and Grafman, J. (1999). "A thoroughly modern gage." *Neurocase*, 5, 345–354.

Doll, B.B., Jacobs, W.J., Sanfey, A.G., and Frank, M.J. (2009). "Instructional control of reinforcement learning: A behavioral and neurocomputational investigation." *Brain Research*, 1299, 74–94.

Doris, J.M. (2002). *Lack of Character: Personality and Moral Behavior*. New York: Cambridge University Press.

Doris, J.M., and Nichols, S. (Forthcoming). "Broadminded: Sociality and the Cognitive Science of Morality." In E. Margolis, R. Samuels, and S. Stich (Eds.), *The Oxford Handbook of Philosophy and Cognitive Science*. Oxford: Oxford University Press

Douglas, M. (1986). *How Institutions Think*. New York: Syracuse University Press.

Doya, K. (2002). "Metalearning and neuromodulation." *Neural Networks*, 15, 495–506.

Doya, K., Ishii, S., Pouget, A., and Rao, R.P.N. (Eds.) (2007). *Bayesian Brain: Probabilistic Approaches to Neural Coding*. Cambridge, MA: MIT Press.

Dreyfus, H. (2002a). "Intelligence without representation: Merleau-Ponty's critique of mental representation." *Phenomenology and the Cognitive Sciences*, 1, 367-383.

Dreyfus, H. (2002b). "Refocusing the Question: Can There Be Skillful Coping without Propositional Representations or Brain Representations?" *Phenomenology and the Cognitive Sciences*, 1, 413–425.

Dunlap, E., Benoit, E., Sifaneck, S.J., Johnson, B.D. (2006). “Social constructions of dependency by blunts smokers: qualitative reports.” *International Journal of Drug Policy*, 17, 171–182.

Ekman, P., and Friesen, W.V. (1971). “Constants across cultures in the face and emotion.” *Journal of Personality and Social Psychology*, 17, 124-129.

Eliasmith, C. (2005). “A New Perspective on Representational Problems.” *Journal of Cognitive Science*, 6, 97-123.

Eliasmith, C. (2003). “Moving beyond metaphors: Understanding the mind for what it is.” *Journal of Philosophy*, C(10), 493-520.

Elster J. (2009). “Social norms and the explanation of behavior.” In P. Hedström, and P. Bearman (Eds.), *The Oxford Handbook of Analytical Sociology*. Oxford: Oxford University Press, 195-217.

Elster, J. (1998). “Emotions and economic theory.” *Journal of Economic Literature*, 36, 47-74.

Elster, J. (1990). “Norms of revenge.” *Ethics*, 100, 862-885.

Elster, J. (1989). “Social norms and economic theory.” *Journal of Economic Perspectives*, 3, 99-117.

Engel, P. (2005). “Tacit Belief,” In W. Østreng (Ed.). *Synergies: Interdisciplinary Communications*. Oslo: Center for Advanced Study, 98-100.

Epstein, J.M. (2001). “Learning to be Thoughtless: Social Norms and Individual Computation.” *Computational Economics*, 18, 9-24.

Ernest-Jones, M., Nettle, D., and Bateson, M. (2011). “Effects of eye images on everyday cooperative behavior: a field experiment.” *Evolution and Human Behavior*, 32, 172-178.

Evans, S., Fleming, S.M., Dolan, R.J., and Averbeck, B.B. (2011). “Effects of emotional preferences on value-based decision-making are mediated by mentalizing and not reward networks.” *Journal of Cognitive Neuroscience*, 23, 2197-2210.

Fehr, E. (2009). “Social preferences and the brain.” In P.W. Glimcher, C. Camerer, R.A. Poldrack, E. Fehr (Eds.). *Neuroeconomics: Decision Making and the Brain*. New York/Amsterdam: Elsevier Academic Press, 215-232.

Fehr, E., and Camerer, C. (2007). “Social Neuroeconomics – The Neural Circuitry of Social Preferences.” *Trends in Cognitive Sciences*, 11, 419-427.

Fehr, E., and Fischbacher, U. (2004). “Third party punishment and social norms.” *Evolution and Human Behavior*, 25, 63-87.

Fehr, E., Fischbacher, U., and Gächter, S. (2002). “Strong reciprocity, human cooperation and the enforcement of social norms.” *Human Nature*, 13, 1–25.

Fehr, E., and Gächter, S. (2002). “Altruistic Punishment in Humans.” *Nature,* 415, 137-140.

Fehr, E., and Schmidt, K.M. (1999). “A theory of fairness, competition, and cooperation.” *The Quarterly Journal of Economics*, 114, 817–868.

Fiorillo, C.D. (2010). “A neurocentric approach to Bayesian inference.” *Nature Reviews Neuroscience*, 11, 605.

Fiser, J., Berkes, B., Orbán, G. and Lengyel, M. (2010). “Statistically optimal perception and learning: from behavior to neural representations.” *Trends in Cognitive Sciences*, 14, 119-130.

Fisher, B., and Tronto, J. (1990). “Toward a Feminist Theory of Caring.” In E. Abel, and M. Nelson (Eds.), *Circles of Care*. Albany: SUNY Press, 36-54.

Fodor, J.A. (1968). “The Appeal to Tacit Knowledge in Psychological Explanation.” *Journal of Philosophy*, 65, 627-640.

Fodor, J.A., and Pylyshyn, Z. (1988). “Connectionism and Cognitive Architecture: A Critical Analysis.” *Cognition*, 28, 3-71.

Forbes, C.E., and Grafman, J. (2010). “The role of the human prefrontal cortex in social cognition and moral judgment.” *Annual Reviews of Neuroscience,* 33, 299-324.

Frank, R.H. (1988). *Passions within reason: The strategic role of the emotions*. New York: W.W. Norton and Company.

Frankfurt, H.G. (2004). *The Reasons of Love*. Princeton, NJ: Princeton University Press.

Frankfurt, H. G. (1982). “The Importance of What We Care About.” *Synthese*, 53, 257-272.

Fredrickson, B.L., and Kahneman, D. (1993). “Duration Neglect in Retrospective Evaluations of Affective Episodes.” *Journal of Personality and Social Psychology*, LXV, 45-55.

Freeman, W.J. (1991). “The physiology of perception.” *Scientific American*, 264, 78–85.

Freeman, W.J., and Skarda, C.A. (1990). "Representations: Who needs them?" In J.L. McGaugh, and N.M. Weinberger & et al. (Eds.), *Brain organization and memory: Cells, systems, and circuits*. London: Oxford University Press, 375-380.

Frey, B.S., Savage, D.A., and Torgler, B. (2010). "Interaction of natural survival instincts and internalized social norms exploring the Titanic and Lusitania disasters." *Proceedings of the National Academy of Sciences of the United States of America*, 107, 4862-4865.

Frijda, N.H. (1986). *The emotions*. Cambridge: Cambridge University Press.

Friston, K. (2010). "The free-energy principle: a unified brain theory?" *Nature Review Neuroscience*, 11, 127-138.

Friston, K. (2009). "The free-energy principle: a rough guide to the brain?" *Trends in Cognitive Sciences*, 13, 293-301.

Friston, K. (2008). "Hierarchical models in the brain." *PLOS Computational Biology*, 4, e1000211.

Friston K. (2005). "A theory of cortical responses." *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, 360, 815-836.

Friston, K., and Stephan, K.E. (2007). "Free-energy and the brain." *Synthese*, 159, 417–458.

Fuster, J.M. (2008). *The Prefrontal Cortex*. (4th edn), London: Academic Press.

Gershman, S.J., and Daw, N.D. (Forthcoming). "Perception, action and utility: the tangled skein." In M. Rabinovich, K. Friston, and P. Varona (Eds.), *Principles of Brain Dynamics: Global State Interactions*. Cambridge, MA: MIT Press.

Gershman, S.J., and Niv, Y. (2010). "Learning latent structure: Carving nature at its joints." *Current Opinion in Neurobiology*, 20, 1-6.

Gibbard, A. (1990). *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Cambridge, MA: Harvard University Press.

Gintis, H. (2010). "Social norms as choreography." *Politics, Philosophy and Economics*, 9, 251-264.

Gintis, H. (2007). "A framework for the unification of the behavioral sciences." *Behavioral and Brain Sciences*, 30, 1–61.

Gintis, H., Bowles, S., Boyd, R., and Fehr, E. (2003). "Explaining altruistic behavior in humans." *Evolution Human Behavior*, 24, 153–172.

Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J.P. (2010). "States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning." *Neuron*, 66, 585–595.

Glenn, A.L., Raine, A., and Schug, R.A. (2009). "The neural correlates of moral decision-making in psychopathy." *Molecular Psychiatry*, 14*,* 5-6.

Glimcher, P.W. (2011). "Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis." *Proceeding of the National Academy of Sciences USA*, 108, 15647–15654.

Goldstein, N., Cialdini, R., and Griskevicius, R. (2008). "A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels." *Journal of Consumer Research*, 35, 472-482.

Gopnik, A., Wellman, H.M., Gelman, S.A., and Meltzoff, A.N. (2010). "A computational foundation for cognitive development: comment on Griffiths *et al.* and McLelland *et al.*" *Trends in Cognitive Sciences*, 14, 342-343.

Grafman, J. (2002). "The structured event complex and the human prefrontal cortex." In D.T.H. Stuss, and R.T. Knight (Eds.), *Principles of Frontal Lobe Function*. Oxford/New York: Oxford University Press, 616.

Griffiths, P. (1997). *What Emotions Really Are: The Problem of Psychological Categories*. Chicago: University of Chicago Press.

Griffiths, T.L., Chater, N., Kemp, C., Perfors, A., and Tenenbaum, J.B. (2010). "Probabilistic models of cognition: Exploring representations and inductive biases." *Trends in Cognitive Sciences*, 14, 357-364.

Hacking, I. (1995). "The looping effects of human kinds." In D. Sperber, D. Premack, and A.J. Premack (Eds.), *Causal Cognition: A Multidisciplinary Debate*. New York: Clarendon Press.

Haley, K.J., and Fessler, D. (2005). "Nobody's watching? Subtle cues affect generosity in an anonymous dictator game." *Evolution and Human Behavior*, 26, 245-256.

Hare, R.D. (2003). "The Psychopathy Checklist—Revised, 2nd Edition." Toronto: Multi-Health Systems.

Harlow, H.F., and Suomi, S.J. (1971). "Social Recovery by Isolation-Reared Monkeys." *Proceedings of the National Academy of Science USA*, 68, 1534-1538.

Harlow, H.F., and Harlow, M. (1962). "Social deprivation in monkeys." *Scientific American*, 207, 136–146.

Haugeland, J. (2002). "Andy Clark on Cognition and Representation." In H. Clapin (Ed.) *Philosophy of Mental Representation*, Oxford: Oxford University.

Haugeland, J. (1998). *Having thought: Essays in the metaphysics of mind*. Cambridge, MA: Harvard University Press.

Herry, C., Bach, D.R., Esposito, F., Di Salle, F., Perrig, W.J., Scheffler, K., Luthi, A., Seifritz, E. (2007). "Processing of temporal unpredictability in human and animal amygdala." *Journal of Neuroscience*, 27, 5958–5966.

Hoffman, E., McCabe, K., and Smith, V.L. (1996). "Social distance and other-regarding behavior in dictator games." *American Economic Review*, 86, 653–660.

Holyoak, K.J. (2008). "Induction as model selection." *Proceedings of the National Academy of Science USA*, 105, 10637-10638.

Houk, J.C. (2007). "Models of basal ganglia." *Scholarpedia*, 2(10):1633.
URL = <http://www.scholarpedia.org/article/Models_of_basal_ganglia>.

Houk, J.C., Adams, J.L., and Barto, A.G. (1995). "A Model of How the Basal Ganglia Generate and Use Neural Signals that Predict Reinforcement." In J.C. Houk, J.L. Davis, D.G. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia*. Cambridge, MA: The MIT Press, 249-270.

Hubel, D.H., and Wiesel, T.N. (1962). "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex." *Journal of Physiology*, 160, 106–154.

Hurlemann, R., Patin, A., Onur, O.A., Cohen, M.X., Baumgartner, T., et al. (2010). "Oxytocin enhances amygdaladependent, socially reinforced learning and emotional empathy in humans." *Journal of Neuroscience*, 30, 4999–5007.

Isen, A.M., and Levin, P.F. (1972). "Effect of feeling good on helping: cookies and kindness." *Journal of Personality and Social Psychology*, 21, 384–388.

Izuma, K., Saito, D.N., and Sadato, N. (2008). "Processing of social and monetary rewards in the human striatum." *Neuron*, 58, 284–294.

Joel, D., Niv, Y., and Ruppin, E. (2002). "Actor-critic models of the basal ganglia: New anatomical and computational perspectives." *Neural Networks*, 15, 535-547.

Johnson, B., Bardhi, F., Sifaneck, S., and Dunlap, E. (2006). "Marijuana argot as subcultural threads." *British Journal of Criminology*, 46, 46–77.

Jones, R.M., Somerville, L.H., Li, J., Ruberry, E.J., Libby, V., Glover, G., Voss, H.U., Ballon, D.J., and Casey, B.J. (2011). “Behavioral and neural properties of social reinforcement learning.” *Journal of Neuroscience*, 31, 13039-13045

Kaas, J.H., and Preuss, T.M. (2008). “Human brain evolution.” In L.R. Squire, D. Berg, F.E. Bloom, S. du Lac, A. Ghosh, and N.C. Spizer (Eds.), *Fundamental Neuroscience*, Third Edition, Amsterdam: Academic Press.

Kawato, M. (2008a). “From "Understanding the brain by creating the brain" towards manipulative neuroscience.” *Philosophical Transactions of the Royal Society B*, 363, 2201-2214.

Kawato, M. (2008b). “Brain controlled robots.” *HFSP Journal*, 2, 136–142.

Kemp, C., and Tenenbaum, J.B. (2009). “Structured statistical models of inductive reasoning.” *Psychological Review*, 116, 20-58.

Kemp, C., and Tenenbaum, J.B. (2008). “The discovery of structural form.” *Proceedings of the National Academy of Sciences USA*, 105, 10687-10692.

Kennett, J. (2002). “Autism, Empathy and Moral Agency.” *The Philosophical Quarterly*, 52, 208, 340-357.

Kim, J. (1998). *Mind in a Physical World*, Cambridge: Cambridge University Press.

Kirsh, D., and Maglio, P. (1992). “Reaction and Reflection in Tetris.” In J. Hendler (Ed.), *Artificial intelligence planning systems: Proceedings of the first annual conference AIPS*. San Mateo, CA: Morgan Kaufmann.

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., and Fernández, G. (2009). “Reinforcement learning signal predicts social conformity.” *Neuron*, 61, 140–151.

Knill, D.C., and Richards, W. (Eds.) (1996). *Perception as Bayesian Inference*. New York: Cambridge University Press.

Knowlton, B.J., and Squire, L.R. (1993). “The learning of categories: Parallel brain systems for item memory and category knowledge.” *Science*, 262, 1747-1749.

Knutson, B., Delgado, M.R., and Phillips, P.E.M. (2009). “Representation of subjective value in the striatum.” In P.W. Glimcher, C. Camerer, R.A. Poldrack, E. Fehr (Eds.). *Neuroeconomics: Decision Making and the Brain*. New York/Amsterdam: Elsevier Academic Press, 389–406.

Knutson, B. (2004). “Sweet revenge?” *Science*, 305, 1246–1247.

Kolodny, J.A. (1994). “Memory processes in classification learning: An investigation of amnesic performance in categorization of dot patterns and artistic styles.” *Psychological Science*, 5, 164-169.

Krueger, F., Barbey, A.K., and Grafman, J. (2009). “The medial prefrontal cortex mediates social event knowledge.” *Trends in Cognitive Sciences*, 13, 103-109.

Kruschke, J.K. (2008a). “Models of categorization.” In R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology*. New York: Cambridge University Press, 267-301.

Kruschke, J.K. (2008b). “Bayesian approaches to associative learning: From passive to active learning.” *Learning & Behavior*, 36, 210-226.

LeDoux, J. E. (2008). “Amygdala.” *Scholarpedia*, 3(4):2698.
URL= <http://www.scholarpedia.org/article/Amygdala>.

Lee, D. (2008). “Game theory and neural basis of social decision making.” *Nature Neuroscience*, 11, 404-409.

Lee, T.S., and Mumford, D. (2003). “Hierarchical Bayesian inference in the visual cortex.” *Journal of the Optical Society of America, A*, 20, 1434-1448.

Lemerise, E.A., and Dodge, K.A. (2008). “The development of anger and hostile interactions.” In M. Lewis, J. M. Haviland-Jones, and L.F. Barrett (Eds.), *Handbook of Emotions, 3rd Ed*. New York: Guilford, 730-742.

Leslie, A.M., Mallon, R., and Dicorcia, J.A. (2006). “Transgressors, victims, and cry babies: Is basic moral judgment spared in autism?” *Social Neuroscience*, 1, 270–283.

Levine, L.J., Safer, M.A., and Lench, H.C. (2006). “Remembering and misremembering emotions.” In L.J. Sanna, and E.C. Chang (Eds.), *Judgments over time: The interplay of thoughts, feelings, and behaviors*. New York: Oxford University Press, 271-290.

Levine, L.J., Prohaska, V., Burgess, S.L., Rice, J.A., and Laulhere, T.M. (2001). “Remembering past emotions: The role of current appraisals.” *Cognition and Emotion*, 15, 393-417.

Lewis, D.K. (1969). *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Li, J., Delgado, M.R., and Phelps, E.A. (2011). “How instructed knowledge modulates the neural systems of reward learning.” *Proceedings of the National Academy of Science USA*, 108, 55–60.

Li, J., Schiller, D., Schoenbaum, G., Phelps, E.A., and Daw, N.D. (2011). "Differential roles of human striatum and amygdala in associative learning." *Nature Neuroscience*, 14, 1250–1252.

Lichtenstein, S., and Slovic, P. (2006). *The Construction of Preference*. New York: Cambridge University Press.

Lin, A., Adolphs, R., and Rangel, A. (2011). "Social and monetary reward learning engage overlapping neural substrates." *Social Cognitive and Affective Neuroscience*.

Lyons, M.J., Akamatsu, S., Kamachi, M., and Gyoba, J. (1998). "Coding Facial Expressions with Gabor Wavelets" *Proceedings, Third IEEE International Conference on Automatic Face and Gesture Recognition*, April 14-16 1998, Nara Japan, IEEE Computer Society, 200-205.

Ma, W.J., Beck, J.M., Latham, P.E., and Pouget, A. (2006). "Bayesian inference with probabilistic population codes." *Nature Neuroscience*, 9, 1432-1438.

Machery, E. (Forthcoming). "Discovery and Confirmation in Evolutionary Psychology." In J.J. Prinz (Ed.), *The Oxford Handbook of the Philosophy of Psychology*. Oxford: Oxford University Press.

Machery, E. (2009). *Doing Without Concepts*. New York: Oxford University Press.

MacKay, D.J.C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge: Cambridge University Press.

Mahajan, N., Martinez, M., Gutierrez, N., Diesendruck, G, Banaji, M., and Santos, L. (2011). "The evolution of intergroup bias: Perception and attitudes in rhesus macaques." *Journal of Personality and Social Psychology*, 100, 387-405.

Marsh, A.A., Ambady, N., and Kleck, R.E. (2005). "The effects of fear and anger facial expressions on approach- and avoidance-related behaviors." *Emotion*, 5, 119-124.

McClelland, J.L., Rumelhart, D.E., and the PDP research group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume II*. Cambridge, MA: MIT Press.

McGeer, V. (2008). "Varieties of Moral Agency: Lessons from autism (and Psychopathy)." In W. Sinnott-Armstrong (Ed.), *Moral psychology, volume 3: The neuroscience of morality*. Cambridge: MIT Press, 227-257.

Medin, D.L., and Schaffer, M.M. (1978). "Context theory of classification learning." *Psychological Review*, 85, 207–238.

Mihov, Y., Mayer, S., Musshoff, F., Maier, W., Kendrick, K.M., and Hurlemann, R. (2010). "Facilitation of learning by social-emotion feedback is beta-noradrenergic-dependent." *Neuropsychologia*, 48, 3168-3172.

Milgram, S., Liberty, H.J., Toledo, R., and Wackenhut, J. (1986). "Response to intrusion into waiting lines." *Journal of Personality and Social Psychology*, 51, 683–689.

Miller, E.K., Freedman, D.J., and Wallis, J.D. (2002). "The prefrontal cortex: Categories, concepts and cognition." *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, 357, 1123–1136.

Milne, E., and Grafman, J. (2001). "Ventromedial prefrontal cortex lesions in humans eliminate implicit gender stereotyping." *Journal Neuroscience*, 21:RC150.

Minsky, M. (1974). *A framework for representing knowledge*. Cambridge, MA: MIT lab memo 306. Excerpts in J. Haugeland (Ed.) (1981). *Mind Design*. Cambridge, MA: MIT Press.

Montague, P.R. (2007). *Your Brain is Almost Perfect: How we make Decisions*. New York: Plume.

Montague P.R, Hyman S.E., and Cohen J.D. (2004). "Computational roles for dopamine in behavioural control." *Nature*, 431, 760–767.

Murphy, G.L. (2002). *The Big Book of Concepts*. Cambridge, MA: MIT Press.

Newell, A., and Simon, H. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nichols, S. (2004). *Sentimental Rules: On the Natural Foundations of Moral Judgment.* New York: Oxford University Press.

Niv, Y. (2009). "Reinforcement Learning in the brain." *Journal of Mathematical Psychology*, 53, 139–154.

Niv, Y., and Schoenbaum, G. (2008). "Dialogues on prediction errors." *Trends in Cognitive Science*, 12, 265–272.

Nosofsky, R.M. (1992). "Exemplar-based approach to relating categorization, identification, and recognition." In F.G. Ashby (Ed.), *Multidimensional models of perception and cognition*. Hillsdale, NJ: Lawrence Erlbaum Associates, 363–393.

O'Doherty, J., Deichmann, R., Critchley, H.D., and Dolan, R.J. (2002). "Neural responses during anticipation of a primary taste reward." *Neuron*, 33, 815–826.

Okasha, S. (2007). "Rational Choice, Risk Aversion and Evolution." *Journal of Philosophy*, CIV, 5, 217-235.

Payne, B.K. (2006). "Weapon bias: Split-second decisions and unintended stereotyping." *Current Directions in Psychological Science,* 15, 287-291.

Pessoa, L., and Adolphs, R. (2010). "Emotion processing and the amygdala: from a 'low road' to 'many roads' of evaluating biological significance." *Nature Reviews Neuroscience*, 11, 773–783.

Pettit, P. (1990). "*Virtus Normativa:* Rational Choice Perspectives." *Ethics*, 100, 725-755.

Piccinini, G. (2007). "Computing Mechanisms." *Philosophy of Science*, 74, 501–526.

Piccinini, G. (2006). "Computational explanation in neuroscience." *Synthese*, 153, 343-353.

Posner, M.I., and Keele, S.W. (1968). "On the genesis of abstract ideas." *Journal of Experimental Psychology*, 77, 353–363.

Preuss, T.M. (2011). "The human brain: rewired and running hot." *Annals of the New York Academy of Science*, 1225 Suppl 1:E, 182-191.

Preuss, T.M. (2009). "The cognitive neuroscience of human uniqueness." In M.S. Gazzaniga (Ed.). *The Cognitive Neurosciences* - Fourth Edition. Cambridge, MA: MIT Press.

Prinz, J.J. (2007). *The Emotional Construction of Morals*. Oxford: Oxford University Press.

Prinz, J.J. (2004). *Gut Reactions: A Perceptual Theory of Emotion*. Oxford: Oxford University Press.

Prior, E., Pargetter, R., and Jackson, F. (1982). "Three Theses About Dispositions." *American Philosophical Quarterly*, 19, 251–257.

Raghanti, M.A., Stimpson, C.D., Marcinkiewicz, J.L., Erwin J.M., Hof, P.R., and Sherwood C.C. (2008). "Cortical dopaminergic innervation among humans, chimpanzees, and macaque monkeys: a comparative study." *Neuroscience*, 155, 203–220.

Raine, A., and Yang, Y. (2006). "Neural foundations to moral reasoning and antisocial behavior." *Social Cognitive and Affective Neuroscience*, 1, 203–213.

Ramsey, W.M. (2007). *Representation Reconsidered*. Cambridge: Cambridge University Press.

Rao, R., Olshausen, B., and Lewicki, M. (Eds.) (2002). *Probabilistic Models of the Brain: Perception and Neural Function*. Cambridge, MA: MIT Press.

Ray, D., King-Casas, B., Montague, P.R., and Dayan, P. (2009). “Bayesian model of behaviour in economic games.” In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (Eds.), *Advances in neural information processing systems*: Vol. 21. MIT Press, 1345–1352.

Ream, G.L., Johnson, B.D., Sifaneck, S.J., Dunlap, E. (2006). “Distinguishing blunts users from joints users: a comparison of marijuana use subcultures.” In S.M. Cole (Ed.), *New Research on Street Drugs*. Hauppauge, NY: Nova Science Publishers, 245-273.

Redgrave, P. (2007). “Basal ganglia.” *Scholarpedia*, 2(6):1825
URL= <http://www.scholarpedia.org/article/Basal_ganglia>.

Reynolds, J.N.J., and Wickens, J.R. (2002). “Dopamine-dependent plasticity of corticostriatal synapses.” *Neural Networks*, 15, 507-521.

Rescorla, R.A., and Wagner, A.R. (1972). “A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement.” In A.H. Black, and W.F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory*. New York: Appleton-Century-Crofts, 64–99.

Rigdon, M., Ishii, K., Watabe, M., and Kitayama, S. (2009). “Minimal Social Cues in the Dictator Game.” *Journal of Economic Psychology*, 30, 358-367.

Robbins, T., and Arnsten, A. (2009). “The neuropsychopharmacology of fronto-executive function: monoaminergic modulation.” *Annual Review of Neuroscience*, 32, 267–287.

Rogers, T., and McClelland, J. (2004). *Semantic Cognition: a Parallel Distributed Processing Approach*, Cambridge, MA: MIT Press.

Rosch, E. (1978). “Principles of Categorization.” In E. Rosch, and B.B. Lloyd (Eds.). *Cognition and Categorization*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Rumelhart, D. E., McClelland, J. L., and the PDP research group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I.* Cambridge, MA: MIT Press

Rushton, J.P., Chrisjohn, R.D., and Fekken, G.C. (1981). “The altruistic personality and the self-report altruism scale.” *Personality and Individual Differences*, 2, 293-302.

Rymer, R. (1993). *Genie: An Abused Child's Flight From Silence.* New York: Harper Collins Publishers.

Sadaghiani, S., Hesselmann, G., Friston, K., and Kleinschmidt, A. (2010). “The relation of ongoing brain activity, evoked neural responses, and cognition.” *Frontiers in Systems Neuroscience*, 23, 4–20.

Sally, D., and Hill, E. (2006). “The development of interpersonal strategy: Autism, theory-of-mind, cooperation, and fairness.” *Journal of Economic Psychology*, 27, 73–97.

Sato, H, and Maharbiz, M.M. (2010). “Recent Developments in the Remote Radio Control of Insect Flight.” *Frontiers in Neuroscience*, 4, 1-12.

Schank, R.C., and Abelson, R.P. (1977). *Scripts, plans, goals and understanding: an inquiry into human knowledge structures*. Hilsdale, NJ: Erlbaum.

Schnall, S., Benton, J., and Harvey, S. (2008). “With a clean conscience: Cleanliness reduces the severity of moral judgment.” *Psychological Science*, 19, 1219-1222.

Schotter, A. (1981). *The economic theory of social institutions*. Cambridge: Cambridge University Press.

Schultz, W. (2007a). “Multiple Dopamine Functions at Different Time Courses.” *Annual Review of Neuroscience*, 30, 259–288.

Schultz, W. (2007b). “Reward.” *Scholarpedia*, 2(3):1652.
URL = <http://www.scholarpedia.org/article/Reward>.

Schultz, W., Dayan, P., and Montague, P.R. (1997). “A neural substrate of prediction and reward.” *Science*, 275, 1593–1599.

Schweighofer, N., and Doya, K. (2003). “Meta-learning in reinforcement learning.” *Neural Networks*, 16, 5-9.

Schwitzgebel, E. (2006/2010). “Belief.” *The Stanford Encyclopedia of Philosophy (Spring 2010 Edition)*, Edward N. Zalta (Ed.), URL= <http://plato.stanford.edu/entries/belief>.

Schwitzgebel, E. (2002). “A phenomenal, dispositional account of belief.” *Nous*, 36, 249–275.

Seamans, J., and Durstewitz, D. (2008). “Dopamine modulation.” *Scholarpedia*, 3(4):2711.
URL = <http://www.scholarpedia.org/article/Dopamine_modulation>.

Searle, J.R. (1995). *The Construction of Social Reality*. New York: Free Press.

Shannon, C. (1948). “A Mathematical Theory of Communication,” *Bell Systems Technical Journal*, 27: 279-423, 623-656.

Shi, L., and Griffiths, T.L. (2009). "Neural implementation of hierarchical Bayesian inference by importance sampling." *Advances in Neural Information Processing Systems*, 22, 1669-1677.

Sienkiewicz-Jarosz, H., Scinska, A., Kuran, W., Ryglewicz, D., Rogowski, A., Wrobel, E., Korkosz, A., Kukwa, A., Kostowski, W., and Bienkowski, P. (2005). "Taste responses in patients with Parkinson's disease." *Journal of Neurology Neurosurgery and Psychiatry*, 76, 40–46.

Sigmund, K., Hauert, C., and Nowak, M.A. (2001). "Reward and punishment." *Proceedings of the National Academy of Sciences USA*, 98, 10757–10762.

Sinnott-Armstrong, W.P. (Ed.) (2008). *Moral Psychology, Volume 2. The Cognitive Science of Morality: Intuition and Diversity*. Cambridge, MA: MIT Press.

Sinnott-Armstrong, W.P. (Ed.) (2008). *Moral Psychology, Volume 3. The Neuroscience of Morality*: *Emotion, Brain Disorders, and Development*. Cambridge, MA: MIT Press.

Smetana, J. (1993). "Understanding of Social Rules," In M. Bennett (Ed.) *The Development of Social Cognition: The Child as Psychologist*. New York: Guilford Press, 111-141.

Smith, A. (1759/1976). *The Theory of Moral Sentiments*. Oxford: Clarendon Press.

Smith, E.E. (2008). "The case for implicit category learning." *Cognitive, Affective, & Behavioral Neuroscience*, 8, 3-16.

Smith, E.E., and Grossman, M. (2008). "Multiple systems of category learning." *Neuroscience and Biobehavioral Reviews*, 32, 249–264.

Smith, J.D. (2002). "Exemplar theory's predicted typicality gradient can be tested and disconfirmed." *Psychological Science*, 13, 437–442.

Smith, M. (2002). "Evaluation, uncertainty, and motivation." *Ethical theory and moral practice*, V, 305–320.

Spitzer, M., Fischbacher, U., Herrnberger, B., Grön, G., Fehr, E. (2007). "The Neural Signature of Social Norm Compliance." *Neuron*, 56, 185-196.

Spreckelmeyer, K.N., Krach, S., Kohls, G., Rademacher, L., Irmak, A., Konrad, K., Kircher, T., and Gründer, G. (2009). "Anticipation of monetary and social reward differently activates mesolimbic brain structures in men and women." *Social Cognitive and Affective neuroscience*, 4, 158–165.

Squire, L.R., and Knowlton, B.J. (1995). “Learning about categories in the absence of memory.” *Proceedings of the National Academy of Sciences USA*, 92, 12470–12474.

Sripada, C. (2005). “Punishment and the strategic structure of moral systems.” *Biology and Philosophy*, 20, 707–789.

Sripada, C., and Stich, S. (2006). “A framework for the psychology of norms.” In P. Carruthers, S. Laurence, and S. Stich (Eds.), *The innate mind: culture and cognition*. Oxford: Oxford University Press, 280–301.

Stich, S. (1993). “Moral Philosophy and Mental Representation.” In M. Hechter, L. Nadel, and R.E. Michod (Eds.), *The Origin of Values*, New York: Aldine de Gruyter, 215–228.

Stone, V.E. (2007). “The evolution of ontogeny and human cognitive uniqueness: Selection for extended brain development in the hominid line.” In S.M. Platek, J.P. Keenan, and T.K. Shackelford (Eds.) *Evolutionary Cognitive Neuroscience,* Cambridge, MA: MIT Press.

Sugden, R. (2002). “Beyond sympathy and empathy: Adam Smith’s concept of fellow-feeling.” *Economics and Philosophy*, 18, 63–87.

Sugden, R. (2000). “The motivating power of expectations.” In J. Nida-Rümelin, and W. Spohn (Eds.), *Rationality, Rules and Structure*. Dordrecht: Kluwer, 103-129.

Sugden, R. (1998). “Normative expectations: the simultaneous evolution of institutions and norms.” In A. Ben-Ner, and L. Putterman (Eds.), *Economics, Values, and Organization*. Cambridge: Cambridge University Press.

Sugden R. (1986). *The Economics of Rights, Cooperation and Welfare*. Oxford: Blackwell.

Suhler, C.L., and Churchland, P.S. (2009). “Control: conscious and otherwise.” *Trends in Cognitive Sciences*, 13, 341-347.

Sunstein, C.R. (2005). “Moral Heuristics.” *Behavioral and Brain Sciences*, 28, 531–573.

Suppes, P. (2007). “Where do Bayesian priors come from?” *Synthese*, 156, 441–471.

Sutton, R.S., and Barto, A.G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Tenenbaum, J.B., Kemp, C., Griffiths, T.L., and Goodman, N.D. (2011). “How to grow a mind: statistics, structure and abstraction.” *Science*, 331, 1279-1285.

Tesauro, G.J. (1995). “Temporal difference learning and TD-Gammon.” *Communications of the ACM*, 38, 58-68.

Tesauro, G.J. (1994). “TD-Gammon, a self-teaching backgammon program, achieves master-level play.” *Neural Computation*, 6/2, 215-219.

Thompson, R.K.R., and Oden, D.L. (1998). “Why monkeys and pigeons, unlike certain apes, cannot reason analogically.” In K. Holyoak, D. Gentner and B. Kokinov (Eds.), *Advances in analogy research: Integration of theory and data from the cognitive, computational, and neural sciences*. Sofia, Bulgaria: New Bulgarian University Press.

Thompson, R.K.R., Oden, D.L., and Boysen, S.T. (1997). “Language-naive chimpanzees (*Pan troglodytes*) judge relations between relations in a conceptual matching-to-sample task.” *Journal of Experimental Psychology: Animal Behavior Processes*, 23, 31–43.

Thornton, C. (2010). “Some puzzles relating to the free-energy principle: Comment on Friston.” *Trends in Cognitive Science*, 14, 53-54.

Tomasello, M., and Call, J. (1997). *Primate Cognition*. New York: Oxford University Press.

Tooby, J. and Cosmides, L. (2008). “The evolutionary psychology of the emotions and their relationship to internal regulatory variables,” In M. Lewis, J.M. Haviland-Jones, and L.F. Barrett (Eds.), *Handbook of Emotions, 3rd Ed.* New York: Guilford, 114-137.

Torrubia, R., Àvila, C., Moltò, J., and Caseras, X. (2001). “The Sensitivity to Punishment and Sensitivity to Reward Questionnaire (SPSRQ) as a measure of Gray’s anxiety and impulsivity dimensions.” *Personality and Individual Differences*, 31, 837-862.

Tricomi, E., Balleine, B., and O’Doherty, J. (2009). “A specific role for posterior dorsolateral striatum in human habit learning.” *European Journal of Neuroscience*, 29, 2225–2232.

Turiel, E. (1983). *The development of social knowledge: Morality and convention,* Cambridge: Cambridge University Press.

Ullmann-Margalit, E. (1977). *The Emergence of Norms.* Oxford: Oxford University Press.

Vilares, I, and Kording, K. (2011). “Bayesian models: the structure of the world, uncertainty, behavior, and the brain.” *Annals of the New York Academy of Sciences*, 1224, 22-39.

von Helmholtz, H. (1925). *Treatise on Physiological Optics, volume III*. Rochester, NY: Optical Society of America.

Walter, H., Abler, B., Ciaramidaro, A., and Erk, S. (2005). “Motivating forces of human actions. Neuroimaging reward and social interaction.” *Brain Research Bulletin*, 67, 368–381.

Weber, E.U., and Johnson, E.J. (2009). “Decisions under uncertainty: psychological, economic and neuroeconomic explanations of risk preference.” In P.W. Glimcher, C. Camerer, R.A. Poldrack, E. Fehr (Eds.). *Neuroeconomics: Decision Making and the Brain*. New York/Amsterdam: Elsevier Academic Press, 127–144.

Whalen, P.J. (1998). “Fear, vigilance, and ambiguity: Initial neuroimaging studies of the human amygdala.” *Current Directions in Psychological Science*, 7, 177–188.

Wolpert, D.M., Doya, K., and Kawato, M. (2003). “A unifying computational framework for motor control and social interaction.” *Philosophical Transactions of the Royal Society of London B Biological Sciences*, 358, 593–602.

Wood, J.N., and Grafman, J. (2003). “Human prefrontal cortex: processing and representational perspectives.” *Nature Review Neuroscience*, 4, 139–147.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Zhong, C., Bohns, V., and Gino, F. (2010). “Good lamps are the best police: Darkness increases dishonesty and self-interested behavior.” *Psychological Science*, 21, 311-314.