Explaining Cognitive Behaviour

A Neurocomputational Perspective

Francesca Micol Rossi

PhD in Philosophy
The University of Edinburgh
2013

To my parents

To Robi

Declaration

I, Francesca Micol Rossi, declare that this thesis is composed by me and that all the work herein is my own, unless explicitly attributed to others. This work has not been submitted for any other degree or professional qualification.

Francesca Micol Rossi

Abstract

While the search for models and explanations of cognitive phenomena is a growing area of research, there is no consensus on what counts as a good explanation in cognitive science.

This Ph.D. thesis offers a philosophical exploration of the different frameworks adopted to explain cognitive behaviour. It then builds on this systematic exploration to offer a new understanding of the explanatory standards employed in the construction and justification of models and modelling frameworks in cognitive science. Sub-goals of the project include a better understanding of some theoretical terms adopted in cognitive science and a deep analysis of the role of representation in explanations of cognitive phenomena. Results of this project can advance the debate on issues in general philosophy of cognitive science and be valuable for guiding future scientific and cognitive research.

In particular, the goals of the thesis are twofold: (i) to provide some necessary desiderata that genuine explanations in cognitive science need to meet; (ii) to identify the framework that is most apt to generate such good explanations.

With reference to the first goal, I claim that a good explanation needs to provide predictions and descriptions of mechanisms. With regards to the second goal, I argue that the neurocomputational framework can meet these two desiderata.

In order to articulate the first claim, I discuss various possible desiderata of good explanations and I motivate why the ability to predict and to identify mechanisms are necessary features of good explanations in cognitive science. In particular, I claim that a good explanation should advance our understanding of the cognitive phenomenon under study, together with providing a clear specification of the components and their interactions that regularly bring the phenomenon about.

I motivate the second claim by examining various frameworks employed to explain cognitive phenomena: the folk-psychological, the anti-representational, the solely subpersonal and the neurocomputational frameworks. I criticise the folk-psychological framework for meeting only the predictive criterion and I stress the inadequacy of its account of cause and causal explanation by engaging with James Woodward's manipulationist theory of causation and causal explanation. By examining the anti-representational framework, I claim that the notion of representation is necessary to predict and to generalise cognitive phenomena. I reach the same conclusion by engaging with William Ramsey (2007) and Jose Luis Bermudez (2003). I then analyse the solely subpersonal framework and I argue that certain personal-level concepts are indeed required to successfully explain cognitive behaviour. Finally, I introduce the neurocomputational framework as more promising than the alternatives in explaining cognitive behaviour. I support this claim by assessing the framework's ability to: (i) meet the two necessary criteria for good explanations; (ii) overcome some of the other frameworks' explanatory limits. In particular, via an analysis of one of its family of models — Bayesian models — I argue that the neurocomputational framework can suggest a more adequate notion of representation, shed new light on the problem of how to bridge personal and subpersonal explanations, successfully meet the prediction criterion (it values predictions as a means to evaluate the goodness of an explanation) and can meet the mechanistic criterion (its model-based methodology opens up the possibility to study the nature of internal and unobservable components of cognitive phenomena).

Acknowledgments

This thesis has greatly benefited from various people.

First of all I would like to thank my Italian supervisors, Edoardo Datteri and Federico Laudisa. They were the first who believed in me and who motivated my move to Edinburgh. They not only encouraged me to do the PhD, but they kept on helping me and advising me on the early drafts of the thesis. Edoardo, in particular, has always been present and has helped me in difficult and non-difficult times.

I am extremely grateful to Andy Clark and to Julian Kiverstein who supported me especially during the first year of my PhD. It hasn't been easy, and they did a great job in putting me on the right track.

I must say a very big thank you to my primary supervisor Mark Sprevak, an exemplary supervisor and an exemplary person. His patience, guidance, advices, punctuality, kindness, precision and attention to details have been inspiring for me. He really taught me how to do philosophy, and, more in general, how to work to high standards!

Many people have helped me working on this thesis. From the philosophy of mind "support group", which has been an extremely useful and friendly place to discuss ideas and find support, to my scout group and staff, to my Italian and Edinburgh-based friends. Special thanks are due to my parents, for their support and for their presence, always, especially in the hard times. And, then, of course, I would like to thank Robi because he has always believed in me, motivated me, helped me and… because, simply, he is Robi. Thanks!

Contents

Introduction

The subject of the thesis can be summarised by the following questions and answers:

Q_1: Which norms and values are used to construct, evaluate and justify models and explanations in cognitive science?

A_1: Currently there are at least four different frameworks that try to explain cognitive phenomena. Each of these frameworks adopts different values and standards.

Q_2: What are the necessary desiderata of a good explanation of cognitive behaviour?

A_2: A good explanation in cognitive science should be predictive and mechanistic.

Q_3: How can we make progress in our understanding of cognitive behaviours?

A_3: Adopting the neurocomputational framework is one way to make progress in our understanding of cognitive behaviours and their underlying processes.

Clarifying these questions and justifying these answers are, in essence, the goals of this thesis.

In this thesis, I will call *cognitive* those behaviours that result from the processing of some kind of information. Examples of cognitive behaviours are: the ability to perceive, to reason, to decide and to remember. Cognitive behaviours are different from reflexes, which are, instead, straight pathways from stimuli to responses that don't require any information processing.

On the one hand, understanding the processes underlying cognition is the primary goal of many disciplines of study, but, on the other hand, there is no consensus on what counts as a good explanation in cognitive science.

It is unclear whether cognitive behaviours need to be explained by adopting a mental or intentional vocabulary, and whether, and to what extent, the dynamics between brain, body and world are required to understand our cognitive life.

There are currently at least four different frameworks that try to explain cognitive behaviour:

1. The *folk-psychological* framework focuses on mental states and their rationalising connections. According to this framework, a cognitive agent's decision to order a glass of water is explained in terms of her desire to have a drink and her belief that water can quench her thirst
2. The *anti-representational* framework. This framework focuses on dynamical loops between brain, body and world. An anti-representational explanation of an animal's wings beating behaviour identifies the dynamic interplay between external feedback from the wings movements (i.e. the couplings between the animal and the environment) and internal regulatory factors as the responsible process for the behaviour
3. The *physiological subpersonal* framework. This framework focuses exclusively on the physical makeup of brains. According to this framework social memory is explained by specifying the roles of a certain neural area and of a specific protein that can be found in it
4. The *neurocomputational* framework. This framework focuses on the informational transactions among neural states. Explaining perception consists in showing how certain Bayesian algorithms are implemented in

specific brain areas and how activities in populations of neurons can transmit relevant information within the cognitive system

It is not clear whether these frameworks are incompatible ways of accounting for cognitive phenomena or whether they (or at least some aspects of them) should be seen as somehow complementary attempts to gain a better understanding of the mind.

What is required for a good explanation of cognitive behaviour in the first place? When is an explanation justified? These are crucial problems that I tackle in this thesis by drawing on the main philosophical positions on the nature of scientific explanation. While answers to the questions above depend, in part, on specific details about the phenomena under study, they also depend in large on the explanatory standards and goals that investigators adopt to determine when explanations succeed and when they fail.

A first goal of the thesis is to examine the various frameworks to make these standards explicit and to show to what extent they depend on different views about the norms governing explanations. By looking at various models at work, the project identifies the norms that the different explanatory frameworks endorse. Results of this analysis provide a better understanding of the nature of the relations among the different frameworks of explanation and of some theoretical terms central in cognitive science, and, in particular, the notion of representation.

A second goal of the thesis is normative. The thesis suggests two necessary features of good explanations of cognitive behaviour: the ability to make predictions and the ability to identify mechanisms. These two criteria are justified by looking at their application to the study of various cognitive capacities. I claim that the neurocomputational framework is the only current framework that can meet both criteria and advance our understanding of cognition. Results of this project are intended to: (i) advance the debate on the explanatory values and standards adopted to construct, evaluate, and justify explanations and models in cognitive science; (ii) guide further research into the nature of cognitive phenomena. The past few years have witnessed an increasing amount of interest by scientists in the distinctive role that Bayesian neurocomputational models play in explaining cognitive phenomena. This has been made possible because of the mathematical advances in identifying predictions from complex probabilistic models. Despite this interest, there is still little philosophical analysis on the neurocomputational framework and its explanatory pay-offs. This thesis is intended to shed light on this framework and on its role in advancing our understanding of cognitive behaviour.

Given the existence of various different perspectives on the study of cognition, it is important to specify what is included in the thesis and what is beyond its scope. The project is not concerned with the localisation of specific functions in the brain. I do not attempt to provide a taxonomy of cognitive versus non-cognitive behaviour. The project is not a conceptual analysis. I do not identify the precise relations that might hold between mental states and brain states. I briefly discuss mental causation, but only in the context of the causal dimensions of psychological explanations. The aim is not to understand how the domain of the mental can be accommodated in that of the physical. The project does not offer a new definition of mechanism. I am also neutral on whether we should be realist or anti-realist about neural mechanisms, whether the component parts of these mechanisms are real parts of the system or artefacts necessary to explain cognitive behaviour. These are all important topics, but they are not directly relevant to the main theme of the thesis. The project is an investigation of the various frameworks and methodologies currently employed to explain cognitive behaviour. My intention is to make their standards and goals explicit and to indicate necessary features of adequate explanations of cognitive phenomena. It is then not the truth of particular models that is defended in this thesis, but the possibility and the suggestion of a fruitful way of achieving adequate cognitive explanations.

Outline of the thesis

The problem that the thesis aims to address concerns what is required for a good explanation in cognitive science.

The goals of the project are twofold: (i) to provide some necessary desiderata that explanations of cognitive behaviour should meet to count as good explanations; (ii) to make the goals and standards adopted by the investigators working within the different frameworks explicit, and to identify the framework that is most apt to generate good explanations.

Chapter 1 introduces the problem and starts to motivate why predictability and identification of mechanisms are two necessary criteria for good explanations in cognitive science by examining the folk-psychological causal framework of explanation.

I argue that folk psychology cannot provide good explanations of cognitive phenomena because it favours predictive power at the expense of mechanisms. In addition to this, I claim that folk psychology offers an inadequate account of cause and causal explanation by engaging with James Woodward's manipulationist theory of causation and causal explanation (e.g. 2003, 2008).

I then conclude that the truth of psychological causal claims cannot be justified by remaining solely at the level of folk psychology, but can be evaluated by descending to the level of mechanism.

Chapter 2 examines the anti-representational framework. In particular, two anti-representational accounts are analysed: Dynamical Systems Theory (e.g. van Gelder, 1995; Chemero, 2000) and Behavioural Systems Theory (e.g. Keijzer, 1998, 2005).

I identify predictability and unification as the main goals of anti-representational explanations. I then claim that they are insufficient to distinguish descriptions from genuine explanations by drawing on general debates on the nature of scientific explanation. I also show that anti-representational explanations are grounded on a weak realisation relation between models and modelled systems, which makes the distinction between genuine explanations and descriptions even more complicated.

I therefore conclude the chapter by arguing that: (i) the identification of localised mechanisms is necessary to complement the predictive power of anti-representational descriptions; (ii) the notion of representation is required to make cognitive behaviour intelligible.

Chapter 3 provides further reasons why representations are necessary to make cognitive behaviour intelligible and to allow generalisations by engaging with William Ramsey's (2007) partial eliminativist claim.

I argue that we can consider a system as trafficking in representations when we explain its cognitive success in terms of internal models that the system employs to draw inferences about the world.

Chapter 4 tackles the so-called "interface problem" (Bermudez, 2005) and analyses the explanatory goals, methodologies and vocabularies of personal- and subpersonal-level explanations of mental phenomena.

I argue that personal-level autonomy theorists' arguments do not succeed because subpersonal information can and sometimes does provide answers to constitutive questions and because a purely normative redescription of a phenomenon runs the risk of being only a hermeneutic but not an explanatory strategy.

At the same time, I claim that purely subpersonal explanations cannot adequately account for cognitive behaviour: (i) certain personal-level concepts are often integral parts of successful explanations of mental phenomena; (ii) folk psychology is not a false theory, but a theory that needs to be enhanced.

I then argue that the methodological autonomy of both personal and subpersonal accounts provides an insufficient starting point to properly explain cognitive phenomena and that both levels are needed.

Chapter 5 discusses Jose Luis Bermudez's tripartite account of rational behaviour (2003) to further argue in favour of mechanistic explanations. In particular, I show that the two criteria for rational behaviour that Bermudez identifies (the behaviour results from a range of alternatives and the behaviour matches some normative standards — i.e. the maximisation of some kind of utility) are inadequate to understand the nature of rational behaviour. Adequate explanations of rational behaviour are only possible when external behavioural criteria of analysis are complemented by internal mechanistic ones: details about how information is encoded and manipulated inside our brains, I claim, are essential to confirm or disconfirm hypotheses about the role and nature of reasoning processes and, ultimately, to evaluate hypotheses about how rationality is naturally possible.

Chapter 6 examines the neurocomputational framework of explanation by analysing the application of Bayesian neurocomputational models to the study of different cognitive behaviours.

I especially focus on various cognitive behaviours that appear to result from some sort of prediction-error minimisation process, which seems to be the main building block of a mechanism that allows agents to perceive what is in the environment, to learn how to predict the consequences of their behaviours, and to perform in a nearly-optimal way.

I then discuss certain experimental data that speak in favour of the existence of some correlation between variables in the models and states in the brain.

Chapter 7 argues that the neurocomputational framework can meet both necessary criteria for good explanations in cognitive science.

In particular, I show that predictions play a central role in neurocomputational explanations and that the framework's openness to an analysis of the possible implementation of cognitive processes together with the growing operationalisation of some of its central claims makes it the most adequate framework to progress in our understanding of various aspects of our cognitive life.

Chapter 1 - The Folk-Psychological Framework

1.1 – Introduction

In this chapter I will begin motivating two necessary desiderata in cognitive science: the ability to predict and to identify mechanisms. In the following chapters I will provide further arguments for this view.

This chapter starts by examining the folk-psychological framework. The folk-psychological framework is typically adopted to make sense of behaviour that cannot be accounted purely in terms of stimulus-response associations. The central idea of the folk-psychological framework is that beliefs, desires and other mental states explain behaviours because they cause them.

I will then engage with James Woodward (e.g. 2003, 2008) who has developed an influential defence of the goodness of folk-psychological explanations. His manipulationist theory of causation and causal explanation can be used to vindicate the truth of folk-psychological causal claims and the adequacy of folk-psychological causal explanations. The main assertion to be considered is that folk-psychological explanations are good and appropriate explanations of cognitive behaviour because they show how a complex event (i.e. a certain combination of beliefs, desires and other mental states) stands in an explanatory relation to a particular behaviour by satisfying certain counterfactuals. According to Woodward, mental states are real causes because they support true counterfactuals and folk-psychological causal explanations are good explanations because they answer *what-if-things-had-been-different* questions. Throughout the chapter I will refer to Woodward's strategy as the "same-level strategy", given that he believes that no further information, beyond that found at the level of the agent, is needed to account for cognitive behaviour.

I attack the view according to which folk-psychological explanations are good explanations of cognitive behaviour by drawing on an analysis of Woodward's same-level strategy. I will argue that: (i) causal claims couched in neural terms can support counterfactuals; (ii) causal explanations couched in neural terms can answer what-if-things-had-been-different questions; (iii) counterfactuals are useful epistemic tools, but they are insufficient to establish the truth of causal claims and the goodness of causal explanations.

At the end of the chapter, I briefly advance the proposal according to which explanations of cognitive capacities don't need to be just predictive, but also mechanistic. I will argue that the identification of mechanisms is necessary to distinguish useful descriptions from good explanations. I will then highlight some pay-offs of lower-level mechanistic explanations with respect to folk-psychological ones and I will defend the thesis according to which the ability to predict is a necessary feature of a good explanation in cognitive science.

1.2 – Explanatory desiderata

A first step to tackle the question of which framework can offer good explanations of cognitive behaviour consists in identifying the necessary desiderata that such explanations need to meet. As I will show throughout the thesis, this question has no obvious answer. There is no agreement in the literature about which features make an explanation a good one. Rather, there are various accounts of explanation, each one stressing some possible candidates.

In this section I will briefly introduce two main families of philosophical theories on the nature of explanation. Although they have been originally elaborated to examine explanations in chemistry and physics, they are useful starting points to understand what explaining cognitive phenomena means.

The first family of philosophical theories is that of the so-called *ontic theories*. According to these theories, a good explanation identifies the real aspects of the world that are in a special relationship with the explanandum phenomenon. This particular relationship is a causal relationship. According to ontic theories, a good explanation identifies the causes of the phenomenon under study (e.g. Salmon, 1984; Woodward, 2003) or its underlying causal mechanism (e.g. Craver, 2007; Bechtel, 2008).

The second broad family of philosophical theories on the nature of scientific explanations includes the so-called *epistemic theories*. According to these theories, a good explanation identifies a special link between what needs to be explained (i.e. the *explanandum*) and what does the explanation (i.e. the *explanans*). Such link has a special epistemic nature: a good explanation of a phenomenon *x* offers information about *x* that is beyond that already provided by the phenomenon *x* itself. In particular, a good explanation provides reasons to expect a certain phenomenon given specific circumstances. The deductive-nomological model of explanation (e.g. Hempel, 1965) advances this idea: to explain a phenomenon is to show that, given certain conditions, the phenomenon is to be expected. Rational expectation is here the mark of a good explanation.

Other epistemic theories have emphasised different desiderata of good explanations, such as the ability of an explanation to unify apparently disparate phenomena (e.g. Friedman, 1974; Kitcher, 1981). According to this view, an explanation is a good explanation when it can show how the explanandum phenomenon fits into a wider framework that can already account for some other familiar phenomena of the world.[1]

This brief description of the current philosophical theories of the nature of explanation shows the lack of consensus on which features make an explanation a good one. The followings are some of the possible desiderata:

- Unification
- Predictability and possibility to control
- Scope
- Identification of causes
- Identification of mechanisms

This lack of consensus, however, doesn't prevent us from identifying those criteria that can, better than others, guide research in methodologically adequate ways. Indeed, we are left with a strong belief that not every methodology or framework of explanation is compatible with our idea of good science.

What are, then, the necessary desiderata of good explanations of cognitive behaviour?

In this chapter I put forward the idea that explanations in cognitive science should be predictive and mechanistic. Explanations should be predictive because good explanations should provide a cognitive advantage, that is, they need to provide information beyond that already offered by the cognitive phenomena themselves. Many philosophers and scientists, with few exceptions[2], indeed agree that predictive power (i.e. the amount of data that corroborates a certain hypothesis) is a necessary,

[1] These accounts are not necessarily exclusive; rather, some features highlighted by one account can also figure in another.

[2] James Woodward (e.g. 2003) and Carl Craver (e.g. 2007) believe that good explanations in the life science do not need to be predictive because it is not possible to identify laws or regular interactions that are responsible for certain cognitive behaviours and that do not admit exceptions, that are not limited in scope and that apply in all times and spaces. For a more detailed discussion on the role of predictions in explanation, see section 1.6.

although not sufficient, feature of good explanations. When we observe that a phenomenon happens quite regularly under certain conditions, we usually tend to explain it by citing those conditions (i.e. whenever conditions X happen, phenomenon Y happens too). Adequate explanations should also be mechanistic, that is, they should identify the components and their interactions that regularly bring the behaviour about. The identification of mechanisms, I will argue, allows to distinguish adequate from inadequate explanations in cognitive science.

1.3 – The folk-psychological framework

The framework that is most commonly employed to explain why humans can perform cognitive behaviour is that of commonsense or folk psychology.

Folk-psychological causal explanations employ certain concepts to make intelligible what someone is doing or did by behaving in a certain way. In particular, they appeal to a class of mental states that are about things, events or states of affairs extrinsic to them, and that figure in explanations of human behaviour. Examples of mental states are beliefs, desires, intentions and expectations. They are often called "propositional attitudes" because it is possible to express them as propositions: belief that something, desire something, and so on. Folk-psychological explanations make reference to these mental states by treating them as causes of behaviour.[3]

[3] Not all folk psychologists believe that mental states are causes and that their connections with behaviour should be understood in causal terms. Some, in particular the so-called autonomy theorists, claim that mental states are *reasons* and that their connections with behaviour are rational connections (e.g. mental states must be consistent and their connections must be governed by familiar deductive principles of logic). I leave this interpretation of folk psychology aside for the moment. I will address it in more details in chapter 4.

A certain behaviour is explained in folk-psychological terms by employing a generalisation of the form “if a person A desires B and believes that by doing C she will get B, then, *ceteris paribus*, she will do C”.

Imagine we want to explain why Sara applied for a job in academia. By employing a folk-psychological explanation, we could say that Sara applied for the job in academia because she desired a job in academia and believed that, by applying for it, she could eventually get it. What makes us able to explain Sara’s behaviour is the fact that we can rely on a generalisation of the form “a person A applies for a job because A wants that job and believes that by applying for it she will get it”.

These kinds of generalisations have a certain degree of success in predicting future behaviour.[4] If we know that Sara desires a job in academia and believes that by applying for jobs in academia she will eventually get one, we can predict that she will apply for them. This prediction can then be confirmed or disconfirmed by Sara’s future behaviour.

A natural move at this point would be to motivate this degree of success in terms of the truth of folk-psychological causal claims: causal explanations involving mental states are often predictive because the causal claims figuring in those explanations are true (e.g. Woodward, 2008). Relatedly, folk-psychological causal explanations are good explanations of cognitive behaviour because they properly capture how we work: humans behave as if they are guided by beliefs, desires and other mental states because they are *actually* guided by them.

In what follows, I will argue that we should resist these conclusions when they are based on folk psychology’s predictive power alone. The reason for this caution has to do with the fact that predictive power is necessary but not sufficient to validate the adequacy of causal explanations. Predictability is a necessary feature of a good explanation because a good explanation should provide us with information about

[4] Not everyone agrees that folk-psychological explanations are highly predictive. Churchland (e.g. 1981), for instance, stresses how folk-psychological explanations often fail to explain and predict many of our cognitive behaviours. I will say more about this in chapter 4.

the explanandum phenomenon that we couldn't have before (i.e. I need to know that, given certain conditions, I should expect the cognitive behaviour). The ability to predict is, however, not a sufficient feature of a good explanation. Recalling briefly the critiques to the deductive-nomological (DN) model of explanation (e.g. Salmon, 1984) will help me justify why predictability is not also a sufficient condition.

According to the DN model of explanation, explaining an outcome is simply a matter of exhibiting nomologically sufficient conditions for it. A good DN explanation has the form of a valid deductive argument that provides one with a rational expectation. Consider the following example:

1. All men who take birth control pills fail to get pregnant
2. Jones is a man who takes birth control pills
3. *Therefore*, Jones fails to get pregnant

This argument is valid (i.e. if the premises were true, the conclusion would be true), but it is not explanatory. The argument is not explanatory because it cites nomologically sufficient but not causally relevant conditions.

A legitimate question arises: given their degree of success, how can we judge whether folk-psychological causal explanations are genuine explanations that pick out the real causes of cognitive behaviour, rather than mere redescriptions of people's behaviour that help us navigate the world?

I will now examine a powerful proposal according to which folk-psychological explanations are indeed good causal explanations and mental states are real causes.

1.4 – Woodward's same-level vindication

One possible way to justify the goodness of folk-psychological causal explanations consists in showing that the causal claims that figure in them support true counterfactuals. The belief that "by applying for an academic job she could get one" caused Sara to apply for the job because, if Sara had believed or desired differently, she would have behaved differently. This form of justification holds that there is nothing more to the truth of the causal claims than the truth of their counterfactuals (e.g. Lewis, 1973).

One of the major advocates of this idea is James Woodward. In *Making things happen* he develops a defence of same-level explanations in terms of counterfactuals by arguing that folk-psychological causal explanations are good explanations because they can answer what-if-things-had-been-different questions. He claims that mental states are genuine causes of cognitive behaviour because we can intervene and manipulate them successfully.

Woodward affirms that many of the arguments typically employed to show that mental states are not causally potent rest on mistaken assumptions about what it is for a relationship to be causal and about what makes an explanation a good causal explanation.[5]

When can we say that X is the cause of Y? According to Woodward, since a cause is something that must make a difference to its effect, X is the cause of Y if and only if were X to be different Y would be different.[6]

[5] Woodward admits that some philosophers draw a clear distinction between providing a causal explanation of a phenomenon and making true claims about the causes of that phenomenon. On his account, however, the two goals are closely related: providing a good causal explanation of a phenomenon requires making true claims about its causes.

[6] The notion of "difference" is here understood in terms of interventions or manipulations, which are, again, causal notions. The fact that Woodward doesn't provide a reductionist account of causation has raised various critiques. An influential one is put forward by Stathis Psillos (2007). He claims that an independent account of the truth-conditions of counterfactuals is required to make manipulationist causal explanations good causal explanations. Consider the following causal claim:

- B_0: X causes Y
- For B_0 to be true, the counterfactual C_1 should be true:
    - o C_1: if X had changed by some intervention I, the value of Y would have changed

The notion of difference and the associated notion of difference-maker are central in Woodward's account. Here, a relation between X and Y is a genuine causal relation if, were an intervention I changing X, the relation between X and Y wouldn't change, while the value of Y would change. A causal relation is then an *invariant* relation. According to this account, it is possible to distinguish causal from merely correlational relations because only the former can be *potentially* exploitable for purposes of manipulation and control.

Consider the relation between attending a private school in the U.S. and scholastic achievement (Woodward, 2008). People tend to believe that students who attend private schools score higher in their final exams than students who attend public schools. Is the attendance at private schools the *cause* of higher scholastic achievement? Is their relation merely correlational? Answers to these, and similar, questions are not obvious. We can imagine various other possible causes of higher scholastic achievement. For example, we could say that parents of private schools' students tend to value more the importance of scholastic achievement and that this directly influences the students' performances or that it is the parents' social-economic status that directly influences scholastic achievements. Let me redescribe the example by means of some variables:

- P: student attending private or public school
- S: measure of the scholastic achievement
- E: parents' social-economic status

- For C_1 to be true, the causal claim B_1 needs to be true:
    - B_1: the intervention I doesn't change the value of Y directly (i.e. by a route independent of X)

"Establishing that certain counterfactuals are true is [then] necessary for establishing that other counterfactuals are true or false." (*ibid.*, p. 101) According to Psillos, counterfactuals provide only an *extrinsic* way of identifying causal relations. Being causal is, instead, an *intrinsic* property of a relation. Accordingly, the truth of causal claims has to be judged on grounds at least partly independent from counterfactual statements.

- A: parents' value of student's scholastic achievement

Within a manipulationist account of causation we can ask whether it is P that causes S by wondering what would happen to S if P were different. Under some interventions on P, if P were the actual cause of S, then S would change. If S does not change under interventions on P, then E or A might be the real cause of S.

Which kinds of interventions or manipulations can we perform on P to judge whether it is the real cause of S? We could run the following experiment. We could divide a group of random students in two sub-groups, and then send one sub-group to attend a private school and the other sub-group to attend a public school. If, independently from the parents' attitude and socio-economic status, the group sent to the private school achieves better results, then we could conclude that P is the cause of S. If this is not the case, we could run other experiments to test the roles of E and A.

To be sure that P is the cause of S, we need to be sure that the intervention occurred on P and not on another variable. Both real and hypothetical interventions can prove the existence of causal relations. Only if interventions are "impossible for (or lack any clear sense because of) logical, conceptual or perhaps metaphysical reasons, then that causal claim is itself illegitimate or ill-defined" (Woodward, 2008, p. 225).

When we conclude that P causes S, we are offering a *type* rather than a *token* causal claim. This means that, within Woodward's manipulationist approach, we can justify the truth of the causal claim "attending a private school causes a higher scholastic achievement", but we might not be able to justify the truth of the claim "Lisa's attendance to Gonzaga private school causes her scholastic achievement to improve".

According to this account of causation based on counterfactuals and on the idea that causes are difference-makers, a good causal explanation can answer what-if-things-had-been-different questions: we can explain an outcome by identifying conditions under which the explanandum-outcome would have been different; these information are about changes and they can be used to manipulate or control the outcome. In Woodward's own words:

> "[…] a successful causal explanation consists in the exhibition of patterns of dependencies (as expressed by interventionist counterfactuals) between the factors cited in the *explanans* and the *explanandum* — factors that are such that changes in them produced by interventions are systematically associated with changes in the explanandum outcome." (*ibid.*, p. 230)

In order to distinguish causally relevant from causally irrelevant (or nomologically sufficient) information, then, we need to assess whether any change in them brings about a change of some sort in the explanandum phenomenon. If a change in the antecedent does not modify the consequent, then the antecedent cannot be considered a relevant cause of the consequent. Accordingly, the main reason why the deductive-nomological (DN) model of explanation doesn't offer a good causal explanation is that it states that certain conditions should figure as causally relevant for an explanandum although they are only nomologically sufficient for it; in other words, deductive-nomological explanations are not good explanations because they cannot answer what-if-things-had-been-different questions.

Consider again the example of Jones discussed above. On Woodward's account, the putative cause is not a real cause because an intervention on it (e.g. Jones doesn't take control pills) doesn't imply any change in the effect (Jones fails to get pregnant). A good causal explanation must, instead, account for counterfactual scenarios.

From the notion of causes as difference-makers and from the claim that there is nothing more to the truth of causal relationships than the truth of the associated counterfactuals, it follows that mental states employed in folk-psychological explanations are the relevant causes of a certain behaviour if their associated counterfactuals are true. Woodward says:

> "[…] all that is required for a change in a mental states M_1 to cause a change in a second mental state M_2 (or in a behaviour B) is that it be true that under some intervention that changes M_1, M_2 (or B) will change. Common sense certainly supposes that episodes like these are very widespread." (*ibid.*, p. 234)

According to Woodward, the folk-psychological framework can provide causally relevant claims, while a reductionist framework often identifies only nomologically sufficient but not causally relevant information.

1.4.1 – Nomological sufficiency and causal relevance

Woodward argues that at the folk-psychological level it is clear that a difference in mental states makes a difference in the outcome behaviour, but that the same cannot be said of lower-level neural activations: a difference in certain neural activations does not always make a difference in the outcome behaviour.

Suppose we want to explain why the pressure of an ideal gas increases from time t_1 to time t_2. We know that at t_1 the gas has temperature T_1, pressure P_1 and that it is in a container with volume V_1. We also know that, after applying heat to the gas, the gas has a new and increased pressure P_2 at time t_2.

One possible explanation, which Woodward calls *macroscopic* explanation, consists in explaining the new pressure by employing the ideal gas law ($PV=nRT$). This law describes how the macro-variables T, V and P relate to each other and change accordingly. By following the law, if the volume of the gas remains stable while the temperature changes, the new pressure P_2 can be explained by the formula $P_2= nRT_2/V$.

We can also explain the new pressure by applying a *microscopic* strategy. We can analyse the molecular configurations and trajectories (G_1) of the gas at time t_1 and its new configuration (G_2) at time t_2. We can then explain P_2 in terms of the force that the new molecular configuration G_2 transfers on the surface of the container.

Woodward argues that, even granting that these micro-level measures are possible, G_2 can't be considered the real cause of P_2: given that there exist many other molecular configurations that could correspond to the same pressure, knowing just one of these configurations is not useful to explain why the gas has pressure P_2 rather than P_3, P_4, and so on.

He applies the same argument to behaviour: we could employ a microscopic strategy to explain, for instance, reaching behaviour by citing the neural correlates of the subject's intention to reach. But then again, given that the same intention could be associated with numerous and different neural configurations, these neural activations are only nomologically sufficient but not causally relevant for the reaching behaviour. The reason for this is that a microscopic explanation in terms of neural activations does not hold true counterfactuals: if the neural configuration had been different, the reaching movement wouldn't have changed.[7] On the contrary, a macroscopic explanation holds true counterfactuals: we can explain a certain reaching behaviour in terms of a specific intention (its cause) such that, if the intention had been different, the grasping behaviour would have been different. According to Woodward, then:

- X causes Y if and only if an intervention changing X would change the value of Y and their relationship would remain invariant
- There is nothing more to the truth of causal claims than the truth of the counterfactuals that hold
- A good causal explanation can answer what-if-things-had-been-different questions and place the explanandum phenomenon into a web of counterfactual dependencies
- A macroscopic causal explanation should be preferred to a microscopic (neural) causal explanation because:

[7] This example refers to some studies carried out by Richard Andersen and colleagues (Musallam *et al.*, 2004). Woodward discusses them as empirical evidences in favour of his account.

- The microscopic causal explanation often contains extremely fine-grained and causally irrelevant nomologically sufficient information
- The microscopic causal explanation can't usually answer what-if-things-had-been-different questions

In the next sections, I will argue that Woodward's same-level strategy is inadequate to vindicate the goodness of folk-psychological causal explanations and the truth of their causal claims. In particular, I will show that not only folk-psychological explanations but also lower-level explanations can answer what-if-things-had-been-different questions and that lower-level causal claims can support counterfactual statements too. In addition to this, I will argue that Woodward's strategy is inadequate to establish the goodness of causal explanations because counterfactual statements are insufficient to justify the truth of causal claims.

In the final section of the chapter I will then briefly argue in favour of lower-level mechanistic explanations that can account for both functional and structural features of components whose regular interactions are responsible for various cognitive phenomena and I will show why good explanations of cognitive behaviour need to be predictive.

1.4.2 – Criticism of Woodward's same-level vindication

It is said that, contrary to the deductive-nomological model of explanation, the relevant cause of a macroscopic behaviour belong to a level where the cause, if changed, would make a difference to its effect: if Sara's belief that by applying for an academic job she will eventually get it were changed, her resulting behaviour would change too.

In this section, I claim that the capacity to support true counterfactuals is indicative of the possible goodness of an explanation, but insufficient to establish it.

Consider the example about the neural correlates of intentions to reach that I discussed above. This example is meant to show that a macroscopic causal explanation in terms of mental states as causes and counterfactuals must be preferred to an explanation that contains only fine-grained details about the nomologically sufficient conditions for a grasping behaviour. Let us look at the example in more detail.

Andersen and colleagues (see Musallam *et al.*, 2004) ran experiments on macaque monkeys to identify the neural correlates of intentions to reach for an object. They recorded the electrical signals of individual neurons in the monkeys' posterior parietal cortex and developed a program to correlate variations in the features of the aggregate firing neurons to differences in intention to reach. Such differences in intention where then observed in the monkeys' overt movements.

The correlations between neural features, associated intentions and consequent reaching movements turned out to be highly predictive: by observing the features of the neural firings, Andersen and colleagues could predict which reaching behaviour would have followed.

Woodward examines these experiments and claims that the identification of the pattern of neural activation (A_1) that corresponds to a specific intention (I_1) to reach for an object (R_1) is not sufficient to conclude that A_1 is the real cause of $R_{1.}$ We are not allowed to conclude this because, according to Woodward, other neural patterns (A_{11}, A_{12}, A_{13}, A_{14}, …) might get activated in other occasions in correspondence to the same intention to reach for the same object. Knowing A_1 can't explain why a monkey performs the behaviour R_1 rather than R_2 (A_1 doesn't inform one about any counterfactual scenarios: were A_1 different, R_1 would still be the same), but it can be nomologically sufficient for R_1.

Despite Woodward's belief that lower-level explanations often provide only nomologically sufficient conditions, a closer look at the methodology adopted in neuroscience reveals that similar experiments are designed to identify repetitive commonalities among neural activations. Once a certain neural pattern is identified as the possible cause of a behaviour, it can be used to test counterfactual scenarios.

Consider Ma and colleagues' study on cue integration (2006). They run different experiments to tackle the following questions:

- How could human perform cue integration of different sensory modalities?
- How could neural activity cause cue integration?

Their aim was to test whether humans performed cue integration by updating the belief about the cause of their sensory input on the basis of sensory information in a Bayesian way. If this were the case, neural activity should encode probabilistic representations of sensory stimuli and integrate them in a Bayesian fashion.

By focusing on the integration of tactile and visual sensory stimuli, they found that cue integration was performed when the activities of cortical neurons, which varied highly from one trial to the next, could be described by Poisson-like statistics. To get clear on what a Poisson-like distribution is, consider the following example. Imagine you normally receive five phone calls each day. There will be days in which you receive less than five phone calls, other days in which you receive more than five phone calls, and days in which you receive none. If we assume that the process responsible for these variations is random, then a Poisson-like distribution tells you how likely it is that you receive a certain number of phone calls during a specific period of time.

Ma and colleagues hypothesised that the presence of Poisson-like variability in certain cortical neurons could enable the brain to carry out Bayesian integration over them.

This study allowed them to generate precise predictions about neural activations and outcome behaviours, and to consider counterfactual scenarios. They could predict that, if the variability of certain cortical neurons were Poisson-like, the subjects' responses would be compatible with those of a Bayesian ideal observer. They could also imagine and test counterfactual scenarios: if the variability of neural activations were not Poisson-like, the performance of the subject in the task would be different from that expected. If, via interventions and manipulations on the activity of certain neurons, they could observe a change in the outcome behaviour, then, by following Woodward's account, a causal explanation of cue integration could cite certain Poisson-like neurons as causes. The explanation would be able to answer what-if-things-had-been-different questions and it would make true claims about cue integration's causes.[8] Woodward himself seems to agree with this conclusion:

> "[…] insofar as the aggregate profile […] of the firing rates that realizes or corresponds to the different ways […] of realizing $I_{1,}$ and [this aggregate] leads to R_1 and [it] contrasts with whatever aggregate profile of neural activity A_2 corresponds to the different intention I_2, it will be equally appropriate to cite A_1 as causing or figuring in the causal explanation for the monkey's exhibiting R_1." (*ibid.*, p. 245–246)

If we were to remain within Woodward's account, we would say that both folk-psychological and lower-level explanations adequately explain cognitive behaviour when they can answer what-if-things-had-been-different questions and support true counterfactuals. For folk psychologists, this conclusion would already be a problem because they would have to provide further reasons for why their explanations should still be preferred to the ones couched in neural terms. However, a more pressing and distinct problem is that Woodward's account cannot be adopted to justify the goodness of any causal explanation given the role he attributes to counterfactual statements.

Counterfactuals are important methodological and epistemic tools to infer the existence of causal relations. They help to "rule out possibilities that are at first promising, or could be thought to be the cause" (Machamer, 2004, p. 31). Saying that

[8] More information is needed in order to conclude that these neurons are the real causes of cue integration. This is due to the fact that counterfactuals are informative but not sufficient to establish the truth of the related causal claims. What we could say, instead, is that neural causes can be explanatory useful even if neurons fire with high variability. Neural causes can be real and genuine causes of human behaviour even if they are somehow different every time.

"had X not occurred, Y wouldn't have occurred either" is important because it informs us about the existence of a relation between X and Y, but the nature of this relation is not expressed via counterfactual claims. There is a "conceptual distinction between causation and invariance-under-intervention: there is an intrinsic feature of a relationship in virtue of which it is causal, an extrinsic symptom of which is its invariance under interventions" (Psillos, 2004, p. 302).

Thinking in terms of counterfactuals is then important to design and run experiments, to develop and modify hypotheses and, ultimately, to discover causal relationships, but counterfactuals alone are insufficient to establish causal connections.[9]

If we reconsider our initial question, by relying on counterfactuals alone, we cannot judge whether we are dealing with a genuine explanation or with a useful redescription of a cognitive behaviour. My suggestion, which I will only briefly consider in the next section and explore more thoroughly in the rest of the thesis, is that in order to understand why folk-psychological explanations have a certain degree of success and whether mental states can cause cognitive behaviour we need to abandon the level of folk psychology and descend to the level of mechanisms. It is the existence of certain underlying physical scattered chains of causal influences that grounds the truth of certain counterfactuals and not vice versa (Clark, 2001).

1.5 – Mechanisms

There is no consensus in the literature concerning what a mechanism really is.[10] I here consider a mechanism to be a set of components whose interactions regularly bring an explanandum phenomenon about.

[9] See section 1.5 for more details on this.

[10] The definition of mechanism I will work with is influenced by Stuart Glennan (e.g. 1996, 2005), Carl Craver (e.g. 2007) and William Bechtel (e.g. 2007, 2008).

What is peculiar of a mechanism and becomes necessary to justify the goodness of an explanation is the fact that its components are identified both functionally (i.e. with reference to the function they play in producing the cognitive behaviour) and structurally (i.e. with reference to certain brain components, their location, shape, size, connections, and so on). This, I argue, consists in identifying what Andy Clark (2001) calls the "real and grounded" causes.

In what follows, I start providing some reasons for why the identification of mechanisms can complement the predictive power of an explanation, thus establishing its adequacy. Consider the following quote:

> "When I claim that some event causes another event, say that my turning the key causes my car to start, I do not believe this simply because I have routinely observed that turning the key is followed by the engine starting. I believe this because I believe that there is a mechanism that connects key-turning to engine-starting. I believe that the key closes a switch which causes the battery to turn the starter motor and so forth. Furthermore, this is not a "secret connexion". I can look under the hood and see how the mechanism works." (Glennan, 1996, p. 50)

Mechanisms are composed of real parts that can be empirically discovered. I am therefore justified in saying "if X hadn't occurred, Y wouldn't have occurred either" if I can point to a mechanism that regularly connects X and Y.

There are various explanatory pay-offs related to the identification of mechanisms underlying cognitive abilities. A mechanism can:

i. Explain why a certain counterfactual holds
ii. Provide further evidence that a system is indeed operating a certain process (i.e. it can distinguish mere descriptions from genuine explanations)
iii. Shed new light on cognitive phenomena

With respect to (i), Woodward claims that the truth-conditions of counterfactuals should not be specified via abstract metaphysical relations of similarity among possible worlds (see Lewis, 1973) or via actual or hypothetical experiments. This, as Psillos (e.g. 2004) notes, leaves open the question concerning the truth-conditions of counterfactual statements.

Psillos explores two possible options. The first option, which Woodward does not endorse, consists in collapsing the truth-conditions of counterfactuals on the evidence-conditions. The consequence of this move would, however, make counterfactuals lose their counterfactuality: they would become similar to future predictions and/or evidences in support of relevant laws. Consider Ohm's law (Psillos, 2004, p. 296) according to which the voltage E of a current is equal to the product of its intensity I times the resistance R of the wire. Take the following counterfactual:

> (C) If the resistance were set to R=r at time t, and the voltage were set to E=e at t, then the intensity I would be i=e/r at t

It t is a future time, then (C) provides an actual conditional, that is, a prediction. If t is, instead, a past time, then, given the existence of good evidence for Ohm's law, (C) provides evidence for the law.

The second option is to provide a story about what these truth-conditions *are* and *how* they are related to evidence-conditions. Psillos' main critique of Woodward is that this story is not present in Woodward's account. This is where mechanisms can enter the picture. As Glennan (e.g. 1996) argues, it is the presence of a mechanism (e.g. thermostat) that explains why a certain counterfactual holds (e.g. if the temperature had risen, the furnace would have turned off) and not vice versa. In other words, the presence of a mechanism linking cause and effect is sufficient to support the truth of certain counterfactual statements:

> "If, for instance, we want to show that smoking causes cancer, the best way to do so would be to discover the mechanism by which tar, nicotine, etc. interact with the body to produce cancerous cells. We might provide overwhelmingly statistical evidence to show the correlation between

> smoking and cancer, but so long as we do not understand the mechanism in question, we can still wonder whether or not the correlation indicates that smoking causes cancer." (Glennan, 1996, p. 66)

When an explanation of a system's capacity must be provided, it is the behaviour of the system that is presented first of all, and not how the behaviour varies or how it would have changed under different conditions. Although it is readily agreed that, when one presents an explanation, one is also committed to a set of counterfactual claims concerning what would have happened if the cause had been different, this is not the same as saying that explanations just consist in exhibiting patterns of counterfactual dependencies. Counterfactual statements play an important role in searching for explanations of cognitive capacities insofar as they help to uncover their mechanisms through experimentations. Not only mechanisms are discovered via experiments, "the rise of mechanical philosophy was closely associated with the rise of experimental science. The observable phenomena of the natural world are to be explained in terms of hidden mechanisms, and these mechanisms are to be inferred using well controlled experiments" (Craver & Darden, 2005, p. 236).

Given that folk-psychological explanations do not aim at providing mechanisms, the existence of certain counterfactual statements is insufficient to validate the truth of their related causal claims and the goodness of their causal explanations.

With respect to (ii), let me briefly introduce an example offered by Gualtiero Piccinini and Carl Craver (2011). They discuss Fodor's position with respect to the explanatory primacy of functional descriptions:

> "If I speak about a device as a "camshaft", I am implicitly identifying it by reference to its physical structure, and so I am committed to the view that it exhibits a characteristic and specifiable decomposition into physical parts. But if I speak of the device as a "valve lifter", I am

> identifying it by reference to its function and I therefore undertake no such commitment." (Fodor, 1968, p. 113)

The description of the valve lifter is a functional description: the valve lifter is a component of an engine that lifts the valve. Given that, from a structural point of view, there can be many different valve lifters (i.e. multiple realisations of the valve lifter), the description of the camshaft and the description of the valve lifter are, according to Fodor, independent from each other. The same argument applies to the generalisations and laws that we find in psychology: these generalisations are not reducible and cannot be captured by those of the lower implementational level.

To argue in favour of a deep relationship between functions and structures, Piccinini and Craver claim that:

> "[…] the "valve lifter" job description puts three mechanistic constraints on explanation: first, there must be valves (a type of structural component) to be lifted; second, lifting (a type of structurally individuated capacity) must be exerted on the valves; and third, there must be valve lifters (another type of component) to do the lifting. For something to be a valve lifter in the relevant respect, it must be able to exert an appropriate physical force on a component with certain structural characteristics in the relevant direction. This is not to say that only camshafts can act as valve lifters. Multiple realizability stands. But it is to say that all valve lifters suitable to be used in an internal combustion engine share certain structural properties with camshafts." (Piccinini & Craver, 2011, pp. 301–302)

Why should an explanation of a cognitive behaviour identify the internal components, their structural properties, their functional capacities and their organisation responsible for a given phenomenon? The answer I put forward here is that the identification of components' structural features offers further evidence that a system is operating a certain process instead of another. This, I argue, is necessary to complement the predictive power of an explanation and to demarcate adequate from inadequate explanations of cognitive behaviour. As Piccinini and Craver clearly point out:

> "[…] if a sub-capacity is a genuinely explanatory part of the whole capacity as opposed to an arbitrary partition (a mere piece or temporal slice), it must be exhibited by specific components or specific configurations of components. In the systems with which psychologists and neuroscientists are concerned, the sub-capacities are not ontologically primitive; they belong to structures and their configurations. The systems have the capacities they have in virtue of their components and organization." (*ibid.*, p. 293).

With respect to (iii), components identified both functionally and structurally appear to shed new light on known phenomena. If we hypothesise, for instance, that certain neurons, characterised spatially and temporally, interact in ways that bring about behaviours that are commonly considered of different types (e.g. motor action and motor imagery), we can test what happens at the neural level in correspondence with both behaviours. If the neural mechanisms that we believe are responsible for motor action are also responsible for motor imagery, then we can say that the two behaviours are related because they are governed by neural components that share important structural, morphological and organisational features. This prediction and the resulting validation would also result in a reconceptualisation of our commonsense thought according to which motor action and motor imagery are two distinct phenomena.

At this lower-level of analysis, we could then have an explanation such that, when the firing is of the type A_1 rather than A_2, we should expect the reaching behaviour B_1 rather than B_2. We could then test our prediction by observing the overt behaviour. The results of this prediction could shed light on novel phenomena (e.g.

certain features of neurons, such as their firing rates, their relations with other neurons and with the resulting behaviour).

Understanding which framework is more suitable to uncover mechanisms is the goal of the current project. The putative framework needs to offer explanations that are both predictive and mechanistic. Achieving this goal requires understanding the relationships between explanations formulated at more or less detailed levels of analysis. This, in turn, yields the inevitable problem of clarifying the relationships between different theoretical notions (i.e. functional and structural/neural), between theories formulated on the basis of different theoretical notions and, more in general, between different disciplines of studies. In chapter 4 I will analyse a version of this problem by making reference to the relationship between personal and subpersonal styles of explanations. I will argue that a co-evolutionary approach (e.g. Churchland, 1986) that favours the integration of information coming from different levels of analysis is required to adequately tackle these problems.

Before concluding this chapter, in the next section I will discuss why good explanations should be predictive.

1.6 – Predictability

Some people (e.g. Woodward, 2003; Craver, 2007) claim that generalisations in the special sciences (e.g. cognitive science, neuroscience and biology) admit exceptions and that the existence of exceptions doesn't affect their explanatory purchase. This thesis undermines the idea, which is central in the deductive-nomological model of explanation, according to which the explanans of a good explanation has to predict the explanandum.

In what follows, I argue in favour of the ability to predict as a necessary, although not sufficient, feature of a good explanation in cognitive science.

Various areas of research in cognitive science aim at offering explanations of cognitive capacities in the form of generalisations that can be empirically tested. Given that these explanations are offered in the form of generalisations, it is plausible to expect that they might not hold for specific organisms, that is, that there could be a number of factors — related to the environment or to the features of the specific organism under consideration — that might cause the failure of these regularities. If we want to explain the behaviour of a specific organism, then it seems that we have to admit that the regularity that we are considering is not exception-free.

It is possible to criticise this conclusion in at least two ways. One way would be to say that it is at least theoretically possible to reformulate the generalisation in a way that it becomes exception-free. We could, for instance, list a number of conditions that have to be present for the generalisation to hold. Another way to avoid the conclusion that regularities admit exceptions would be to add to the generalisation a *ceteris paribus* clause (i.e. "in normal circumstances").

Despite the difficulties of both strategies[11], some believe that it is not even necessary to avoid exceptions. Woodward argues that a generalisation that admits exceptions can still be explanatory. By accepting this position, we then need to abandon the idea that a generalisation has to allow the prediction of the corresponding explanandum behaviour.

Craver strongly supports this thesis with respect to mechanistic explanations in biology. He claims:

> "[…] explaining a phenomenon need not require showing that it was to be expected. […] In neuroscience (and, in fact, also in physics, chemistry, and almost everywhere) improbable things happen, and when

[11] It is often claimed that adding *ceteris paribus* clauses, which is a strategy that is adopted in various areas of scientific research (see Earman *et al.*, 2002), yields vacuous and non-empirically testable claims: the content of "*ceteris paribus*, all Fs are Gs" can be interpreted as "all Fs are Gs, unless the contrary holds".

> they do, mechanisms can explain them as well […].” (Craver, 2007, p. 39)

It is important to stress that the thesis according to which the ability to predict is not a necessary feature of a good explanation is distinct from the thesis according to which the ability to predict is not a sufficient criterion for a good explanation. As I have described in section 1.3, a derivation from premises to conclusions can satisfy the criteria of the deductive-nomological model of explanation and yet not yield an adequate explanation because it cites only nomologically sufficient but not causally relevant conditions with respect to the explanandum phenomenon. The claim I am examining here is different: the ability to predict is not even a necessary desideratum of good explanations. Both Woodward and Craver affirm that a good explanation doesn’t have to offer reasons to expect the explanandum phenomenon.

Consider the behaviour of place cells (O’Keefe & Conway, 1978) discussed by Edoardo Datteri and Federico Laudisa (2012). In searching for the possible mechanisms underlying spatial memory in rats, in 1970s investigators found that certain cells in the area CA1 of the hippocampus of rats fire whenever the animal moves through a particular location in space. The generalisation is, in this case, the following:

> (G) Each place cell fires only if the rat is located at a particular spatial position

This generalisation is considered explanatory to the extent that it identifies the features that generate the relevant behaviour. Given that the generalisation (G) is considered explanatory, (G) is used by investigators to predict the behaviour of place cells and to test the predictions on the basis of experimental results. If there are discrepancies between predictions and data, investigators are in a position to reconsider the adequacy of the generalisation that motivated those predictions. Denying that generalisations should have predictive power means, somehow, that generalisations are not even explanatory.

Another problem related to the claim that predictability is not a necessary desideratum of a good explanation concerns the empirical control of the hypothesis. In Datteri and Laudisa's own words:

> "This generalization [G] gives rise to a prima facie testability issue that has frequently been discussed in the philosophical literature (Earman et al., 2002). Suppose we are able to monitor the activity of a rat place cell, pc, while the animal is running in its environment. We are free to make predictions about the rat based on [G]: for example, we can predict that the next time pc fires, the rat will be in the vicinity of position <x,y> (where <x,y> is the centre of the receptive field). However, given the assumption that [G] is conditional to the absence of several perturbing factors, and if we have no idea of what these perturbing factors are, such a prediction is not much worth betting on: the behaviour of pc in "real-world" settings could well be perturbed by some unknown factor, and the prediction is likely to fail." (Datteri & Laudisa, 2012, p. 603)

Affirming that an explanation can admit exceptions and that it doesn't have to be predictive, then, generates testability problems that depend on the conventional methodology to compare predictions with experimental results. When should prediction failures count against the hypothesised generalisation? If we follow Craver and Woodward and, more in general, if we believe that adequate explanations of behaviour don't need to be predictive, we don't have a straightforward answer to this question. Discrepancies between predictions and data might be due to the fact that the generalisation itself is incorrect or they might depend on the presence of some unknown perturbing condition in the experimental setting.

Let us reconsider for a moment Woodward's criteria for identifying good explanations. According to Woodward, explaining is a matter of showing patterns of counterfactual dependencies; a good explanation is able to answer what-if-things-had-been-different questions and a causal relation is an invariant relation.[12] This

[12] For more details on Woodward's account, see section 1.4.

means that the ability to predict *is* central in Woodward's manipulationist model of explanation. However, he explicitly criticises the idea that predictability is a necessary feature of good explanations and affirms that explanatory generalisations can admit exceptions. He discusses two examples to argue in favour of this idea. The first example concerns the relationship between untreated syphilis and paresis. Given that only 25% of people with syphilis get paresis, the existence of untreated syphilis doesn't allow us to predict the development of paresis. Nevertheless, Woodward argues that the following is a good explanation of why Jones has paresis:

> Jones' paresis is caused by his untreated syphilis

The second example concerns a subject hitting the edge of a table with her knee, thus turning over an ink-bottle whose content ruins the carpet. The following claim is considered an adequate explanation of the fact that the ink-bottle turned over:

> Knocking over the table with the knee caused the ink-bottle to turn over and ruin the carpet

In both examples, the premises of the explanations (i.e. Jones' untreated syphilis and knocking over the table) don't provide the basis to predict the conclusions (i.e. Jones's paresis and ink-bottle turned over) in the absence of other circumstances. Nevertheless, according to Woodward, both explanations are adequate explanations, hence good explanations don't have to predict the explanandum phenomenon.

This claim seems to be in contrast with Woodward's account according to which a good explanation should answer what-if-things-had-been-different questions. By answering these questions, an adequate explanation provides ways to control and intervene on the explanandum phenomenon. If the correct explanation of Jones' paresis is his untreated syphilis, then one should be in a position to answer the following question: if Jones had had (or hadn't had) latent syphilis, would have he developed paresis? It doesn't seem that the above explanation can provide an answer to this question. At the same time, were the explanation to be adequate, one should be able to intervene on the cause (Jones' syphilis) to modify the effect (Jones' paresis). The above explanation, however, doesn't provide us with adequate strategies of control and intervention. In particular, the explanation doesn't allow us to prevent other people with untreated syphilis to develop paresis. Nevertheless, being able to control a phenomenon by knowing the conditions under which it is to be expected is an important pay-off of good explanations: adequate explanations should allow us to know that, under certain conditions, a phenomenon is to be expected. Knowing this, one can work out ways to modify the conditions so that the phenomenon doesn't come about. For all these reasons, it is difficult to see how the claims above can actually figure as good explanations with respect to Woodward's model of explanation.

Along the same lines, Craver argues that a good explanation should identify a mechanism, which is composed of parts that causally interact to bring the explanandum phenomenon about. Craver's notion of causality is borrowed from Woodward's account: a casual relation is an invariant relation under intervention. The description of a mechanism, then, allows us to answer what-if-things-had-been-different questions.

According to Craver, in order to explain "[…] one needs to know how the phenomenon changes under a variety of interventions into the parts *and* how the parts change when one intervenes to change the phenomenon" (Craver, 2007, p. 160). It is therefore difficult to justify Craver's claim that "explaining a phenomenon need not require showing that it was to be expected". On the contrary, he seems to be saying that a good explanation identifies the conditions — the parts and their activities — under which a certain phenomenon is to be expected. If these conditions are known, we can intervene on them to change the explanandum phenomenon.

It is therefore reasonable to conclude that the ability to predict is central to both Woodward's and Craver's accounts. More generally, the necessity to predict is required to study behaviours and phenomena. Denying the necessary role of predictions generates, as I have just shown, important testability problems, which, in turn, could yield explanatory failures or inadequacies.

1.7 – Conclusion

In this chapter I began to tackle the question concerning the necessary desiderata of good explanations of cognitive behaviour. I suggested that the ability to predict and the ability to identify mechanisms are two necessary desiderata of good explanations in cognitive science.

I argued that the folk-psychological framework is ill-suited to generate good explanations because it favours predictive power at the expense of mechanisms. I claimed, contrary to Woodward, that counterfactuals statements are insufficient to validate the truth of causal claims and the goodness of causal explanations: counterfactuals are important epistemic tools insofar as they can help to uncover mechanisms.

I argued that the truth or falsity of folk-psychological causal claims cannot be justified by remaining at the level of folk psychology, but that it can be evaluated by descending to the level of mechanisms. I then briefly addressed the explanatory pay-offs of mechanistic and predictive explanations that go beyond their ability to distinguish genuine explanations from mere descriptions of cognitive phenomena.

Chapter 2 - The Anti-Representational Framework

2.1 – Introduction

In the previous chapter, I introduced the ability to predict and to identify mechanisms as two necessary desiderata of good explanations in cognitive science and I argued that the folk-psychological framework cannot provide such explanations mostly because it favours predictive power at the expense of identification of mechanisms. In this chapter I will examine the so-called anti-representational framework. This framework aims at explaining cognitive behaviour without invoking the notion of representation.

In what follows, I will analyse two anti-representational accounts: Dynamical Systems Theory and Behavioural Systems Theory. I will show that the anti-representational framework is ill-suited to properly explain cognitive behaviour and I will claim that: (i) the identification of localised mechanisms is necessary to complement the predictive power of Dynamical Systems Theory's formal descriptions; (ii) the notion of representation is important to make cognitive behaviour intelligible.

2.2 – Dynamical Systems Theory

According to Dynamical Systems Theory (DST), explanations of cognitive behaviour don't require the appeal to the notion of representation. Abandoning the notion of representation is seen as a first necessary step to radically change the way in which cognitive scientists think and study how the brain carries out cognitive tasks.

Dynamicists (e.g. van Gelder, 1995; Chemero, 2000) put forward various reasons why we should reject "sophisticated internal representations" (van Gelder, 1995, p. 346) and embrace a radically new framework of explanation. Their basic complaint is that physical systems can engage in many or all of the various cognitive tasks for which cognitivists have postulated internal representations without employing internal representations.

A prototypical example of a system performing a cognitive task, according to dynamicists, is the Watt's centrifugal governor for the steam engine (*ibid.*). Watt designed the governor to solve the problem of maintaining constant speed for the flywheel of a steam engine. The governor consists of a vertical spindle attached to a flywheel that rotates with a speed proportional to the speed of the flywheel. Two arms with metal balls on their ends are attached to the spindle and are free to move with a force that is proportional to that of the speed of the governor. Thanks to a mechanical device, the angle of the arms changes the opening of a valve, thus controlling the amount of steam driving the flywheel. If the flywheel turns too fast, the arms will rise and the valve will partially close. This closure will then reduce the amount of steam available to turn the flywheel, thereby slowing the flywheel down. If, instead, the flywheel turns too slowly, the arms will drop, thus causing the valve to open. This opening will make more steam available, hence it will allow the speed of the flywheel to increase.

Advocates of DST typically employ the Watt governor example to show that cognitive tasks can be carried out without employing internal representations. Their central claim is that the Watt governor can solve problems for which people might be tempted to posit a representational solution (e.g. we might be tempted to interpret the present and desired speed of the flywheel and/or the opening and closing of the valve in representational terms) without processing any representation. van Gelder offers various arguments to show why we should resist this temptation.

The first argument consists in noticing that the existence of causal relationships between certain parts of the governor (e.g. the causal connections between the arm angle and the engine speed) is not sufficient to conclude that the former is a representation of the latter (e.g. that the arm angle represents the engine speed).

The second argument has to do with the explanatory pay-offs that a representational story, if needed, should provide. van Gelder argues that treating the governor as a device that manipulates representations doesn’t provide any specific explanatory pay-off.

The third argument concerns the notion of representation itself. Advocates of DST claim that the notion of representation is not rich enough to account for the highly complex dynamics that exist between the arm angle and the engine speed. Here the idea is that the governor and its environment (the steam engine) are so closely linked that considering one of them as representing the other wouldn’t give an adequate explanatory purchase with respect to the behaviour of the system in question.

Port and van Gelder extend the above arguments to all kinds of cognitive tasks. They argue that cognitive behaviour can be adequately explained in purely DST non-representational terms because explaining a cognitive behaviour is a matter of identifying its relevant parameters and their coupled dynamics in the course of the physical system’s evolution over time:

> “The cognitive system is not a discrete sequential manipulator of static representational structures; rather, it is a structure of mutually and simultaneously influencing changes. The cognitive system does not interact with other aspects of the world by passing messages or

> commands; rather, it continuously coevolves with them. [...] To see that there is a dynamical approach is to see a new way of conceptually reorganizing cognitive science as it is currently practiced." (Port & van Gelder, 1995, p. 24)

Cognitive systems can, therefore, be adequately explained, according to DST, by means of mathematical descriptions in terms of feedback loops. These formal descriptions can predict how cognitive behaviours unfold over time.

To better understand the explanatory pay-offs of DST, let me consider its main features and goals in more detail.

2.2.1 – Mathematical descriptions

A mathematical description of a dynamical system, which is called *system's state,* is a formal description that consists in a n-dimensional mathematical space, whose dimensions correspond to the state variables of the system. These state variables are measured quantities that supervene on the behaviour of the system's lower constituents and that do not correspond to any particular part of the system. State variables are often called *lumped parameters*, that is, parameters that are equal or proportional to some average value of the corresponding distributed ones. The mathematics employed typically specifies a dynamical law that determines how the values of the state variables evolve through time.

According to a dynamical systems analysis, cognition can be explained as a multi-dimensional space of all possible thoughts and behaviours, which is traversed by a path of thinking, where certain environmental and internal pressures, which are captured by mathematical equations, influence the path that a subject follows in the space.

The set of all trajectories is called the *flow* and its features are the objects of study of DST. To help determine the shape of the flow, DST relies on a number of constructs, which include the notion of "attractor" (i.e. a point or a region in the state space where all the trajectories that pass close to it get sucked into it), the notion of "basin of attraction" (i.e. the area of influence of an attractor), and "bifurcations" (i.e. points in the system's state where a small change in the state values can modify the flow).

Thanks to this mathematical apparatus, which doesn't posit any role for representations, DST is said to adequately explain all phenomena that unfold over time, cognitive phenomena included.

Consider the case of coordinated finger movements. The dynamical HKB model (Kelso, 1995) — named after its originators (Haken, Kelso and Bunz, 1985) — has become a paradigmatic example of a successful application of DST tools. The HKB model is based on the observation that, when asked to oscillate with the same frequency both index fingers back and forth, people produce only two basic patterns. One pattern consists in both fingers moving to the left or to the right at the same time. This pattern is called "in-phase" motion. In the second pattern, one finger moves to the left and the other finger moves to the right. This is called "anti-phase" motion.

The HKB model characterises the temporal evolution of one purely behavioural variable (i.e. the relative phase of the fingers) as a function of another purely behavioural variable or *control parameter* (i.e. the fingers' oscillation frequency). Both variables do not correspond to any internal state that *represents* the frequency of the movements or the state of the fingers' coordination with respect to one another. Interestingly, if people in anti-phase motion are asked to increase the frequency of oscillation, they will spontaneously switch to the in-phase mode at a certain frequency of movement (the so-called "critical region"). If, instead, people start with in-phase motion, they won't exhibit such switch. In other words, the in-phase motion will remain stable through and beyond the critical region. In the language of dynamics, there are two stable attractors at low frequencies and a bifurcation at a critical point, leading to only one stable attractor at higher frequencies.

The features of the HKB model can be captured using the mathematics of Dynamical Systems Theory. The model describes how one collective variable (i.e. relative phase) varies depending on the control parameter (i.e. frequency of oscillation). The variable ϕ in the model is an abstract mathematical magnitude that corresponds to the oscillation phase, which is instead a concrete behavioural quantity. The coordination law can be expressed as:

$$\phi = -\sin\phi - 2k\sin 2\phi$$

The parameter k in the model corresponds to the inverse of the oscillation frequency in the experiment, such that an increase in frequency corresponds to a decrease in k.

It is important to note two features of HKB model. First, the model accounts for the data without positing any kind of "inner switching mechanism"; rather, the switch results from the self-organising evolution of the system. Second, the model makes novel predictions that were unknown at the time the model was developed. The model can, for instance, predict the consequences of selective interference by applying an electrical pulse to the subject's hand so as to disrupt the normal coordination of movements.

Interestingly, the HKB model works because there is a correspondence between the quantitative properties of an abstract variable and the quantitative properties of its concrete counterpart. This is to say that the concrete system *instantiates* the mathematical system.

2.2.2 – Unification

The importance that dynamicists attribute to mathematical descriptions doesn't depend only on the fact that formal descriptions are useful tools for explaining behaviour that unfolds over time, but also on their ability to shed light on the real nature of dynamical processes, as it has already happened in other sciences, primary in physics. By providing dynamic explanations, advocates of DST can unify multiple phenomena under a common generalisation. The dynamical equations of the HKB model, for instance, can be used to describe similar coordination patterns implemented across physically disparate systems.[13] As Chemero claims:

> "HKB model is an example of a general strategy for describing constraints on behavior. First, observe patterns of macroscopic behavior; then seek collective variables (like relative phase) and control parameters (like rate) that govern the behavior; finally, search for the simplest mathematical function that accounts for the behavior. Because, HKB argue, complex systems (like the one involving the muscles, portions of the central nervous system, ears, and metronome in the finger-wagging task) have a tendency to behave like much simpler systems, one will often be able to model these systems in terms of extremely simple functions, with only few easily observable parameters, which reflect the dynamic behavior." (Chemero, 2001, p. 141)

Dynamicists believe that the use of mathematics is necessary not only to unify disparate phenomena, but also to naturalise cognition (e.g. Chemero, 2000; Swenson & Turvey, 1991; Turvey & Carello, 1981): if we can understand cognition by using the same method and mathematical apparatus that we already use in other sciences, we will be in a position to offer a proper naturalised account of cognitive behaviour.

2.2.3 – Dynamics

The use of mathematical equations for explaining cognitive phenomena also brings to the foreground the importance of real-time dynamics: cognition is active and cognitive processes are influenced by time and by changes in the environment.

[13] Coordination patterns that can be modelled with HKB are: certain aspects of motor skill learning, interpersonal coordination, speech perception and visual perception (see Kelso, 1995).

The primary role attributed to real-time dynamics is an important novelty of the DST approach that puts it immediately in contrast with other theoretical frameworks. The natural competitor is the Classical Computational Theory of Mind that considers cognitive tasks as the result of processes operating on discrete symbols in atemporal manner.[14] According to DST, adequate explanations of cognitive behaviour need to consider cognition in its complexity, as strongly dependent on time and system-environment interactions. This is the so-called strong coupling thesis that leads up to anti-representationalism: cognitive agents are embedded in an environment with which they strongly interact. This means that the states of a system that are responsible for its cognitive activity are dynamically coupled with external environmental states and features, with the consequence that any change in the former brings about, simultaneously, changes in the latter. At the same time, the affected states of the environment influence the change in the states of the system. These influences are sometimes called closed-loop control processes.

Given these close couplings and influences between the system and its environment, only DST tools, and not representational or computational tools, are considered adequate to explain cognition.

To sum up, according to DST advocates:

- Mathematical descriptions are:
    - The most suitable tools for analysing cognition and for predicting cognitive behaviour unfolding over time
    - Sufficient for explanations, hence there is no need to state the existence of internal representations
    - Successful strategies for understanding dynamical systems in other sciences, so they might offer a good methodology to arrive at a naturalised account of cognition

[14] For a discussion on the Classical Computational Theory of Mind, see chapter 3.

- Cognition needs to be considered in its complexity, as a phenomenon that unfolds over time and that is influenced by system-environment interactions

2.3 – Critical discussion

2.3.1 – The explanatory inadequacy of formal mathematical methods

A DST description of a cognitive process is a formal mathematical description whose main building blocks are the system's variables. These variables constitute the state space of the system and are called *lumped parameters* because they do not correspond to any particular constituent of the system.

Given that it would be impossible to consider all the variables involved in the generation of even a very small reflex, the choice of which variable should be included in the formal mathematical model is crucial to correctly model the explanandum phenomenon.

Within a DST model, this important choice is made independently from the substrate that realises the explanandum cognitive process. Indeed, as I have claimed above, implementational details are not seen as necessary for good explanations.

In this section, I defend the claim according to which predictability is the main goal of dynamical systems explanations. In Port and van Gelder's own words: a dynamical model "yields not only precise descriptions […] but also predictions which can be used to evaluate the model" (Port & van Gelder, 1995, p. 15), and for a dynamical model to be predictively adequate, it doesn't necessary have to include implementation-dependent variables. Let me elaborate on this point.

In his 1998 paper, van Gelder considers various possible objections to DST. One of these objections concerns the distinction between description and explanation: can DST provide genuine explanations of behaviour? The thought behind this objection is that, given any set of data, one can construct a line that connects them, and any line can be approximately described by some equation or another. The risk for DST is that of "curve fitting", that is, connecting the data, formulating an equation that describes the connecting line, and then seeing the result as the explanation of the phenomenon that generated the data. van Gelder himself admits that:

> "A *poor* dynamical account may amount to little more than ad hoc 'curve fitting', and would indeed count as mere description." (van Gelder, 1998, p. 625)

According to van Gelder, the fact that some dynamical accounts are poor doesn't depend on them being dynamical. Indeed, van Gelder notes that genuine explanations similar to the dynamical ones are found in many other sciences. In addition to this, DST explanations are not mere descriptions because they can formulate novel predictions and support counterfactuals (see Clark, 1997). Consider the following quotes:

> "Dynamical modelling […] involves finding […] a mathematical rule, such that the phenomena of interest unfold in exactly the way described by the rule." (Port & van Gelder, 1995, p. 14)

> "[…] taking some novel phenomena and showing that it is the behavior of a dynamical system is always a significant scientific achievement." (*ibid.*, p. 11)

> "Such models specify how change in state variables at any instant depends on the current values of those variables themselves and other parameters. Solutions to the governing equations tell you the state that the system will be in at any point in time, as long as the starting state and the amount of elapsed time are known." (*ibid.*, p. 19)

According to DST, then, cognitive phenomena are explained by citing the laws (e.g. differential equations) and certain initial conditions that govern the target systems.

The central role of predictions in DST explanations and the lack of logical distinction between an explanation of a given state and its prediction make DST explanations a case of deductive-nomological explanations.

In line with the claims I made in the previous chapter, predictive success, in this case the predictive success of dynamical models, is not sufficient to provide good explanations of cognitive phenomena. While dynamicists believe that we can dispense from decomposition and localisation (Bechtel, 1998) and still obtain an adequate explanation of a cognitive capacity, I argue that information concerning the nature and the role of the internal physical components responsible for a certain behaviour are necessary to explain it. In particular, the identification of the responsible mechanism offers strategies of control and testability that allow us to distinguish a mere description from a proper explanation. This, in turn, requires a deep analysis of the material substrate that brings about the cognitive process.

A mechanistic approach is not only necessary to distinguish predictive descriptions from genuine explanations, but it is also needed to avoid problems related to certain dynamical claims. Consider the following quote:

> "[…] the relationship at the heart of the nature hypothesis [i.e. the hypothesis that tells us what cognitive agents are by specifying the relation they bear to dynamical systems] is not identity but *instantiation*. Cognitive agents are not themselves systems (sets of variables), but, rather, objects whose properties can form systems. Cognitive agents instantiate numerous systems at any given time. According to the nature hypothesis, the systems responsible for cognitive performance are dynamical. […] Another noteworthy fact about these models is that the variables they posit are not low level (e.g. neural firing rates), but, rather, macroscopic quantities at roughly the level of the cognitive performance itself." (van Gelder, 1998, p. 619)

A cognitive system is, then, a real dynamical system when it changes over time and when it instantiates a mathematical dynamical system that correctly describes some aspects of its change:

> "[…] for every kind of cognitive performance exhibited by a natural cognitive agent, there is some quantitative system instantiated by the agent at the highest relevant level of causal organization, so that performances of that kind are behaviors of that system: in addition, causal organization can and should be understood by producing dynamical models, using the theoretical resources of dynamics, and adopting a broadly dynamical perspective." (*ibid.*, p. 622)

How should we interpret the claim that "all cognitive systems are dynamical systems"? Marco Giunti (1995) suggests that we could read the claim as saying that all dynamical systems are real dynamical systems, that is, systems that change over time. Such interpretation, however, would yield a trivial thesis given that any concrete object can be said to change over time, in some sense. We could instead interpret the claim as saying that all dynamical systems (e.g. mathematical dynamical systems) are real dynamical systems, in which case the thesis will be an absurd: a cognitive system, which is a real object, cannot be identical to a mathematical dynamical system, which is, instead, an abstract formal structure. A third reading, Giunti says, could make more sense. It would interpret the claim as "all cognitive systems instantiate dynamical systems", which means that the study of mathematical dynamical systems can help us to understand something about cognitive systems. The specific sense of "instantiation" becomes clear in van Gelder's own words:

> "The scientist furnishes an abstract dynamical system to serve as a model by specifying abstract variables and governing equations. Simple models can be fully understood by means of purely mathematical techniques. More commonly, however, scientists enlist the aid of digital computers to *simulate* the model (i.e. compute approximate descriptions of its

> behavior). The simulation results are compared to experimental data from the target. To the extent that the correspondence is close, the target system is taken to be similar in structure to the abstract dynamical model." (van Gelder, 1998, p. 620)

A mathematical dynamical model, then, allows to simulate certain aspects of the behaviour of a cognitive system by first implementing the model and then assigning to the model a task similar to the one assigned to the cognitive system. In carrying out the task, the simulated model goes through a changing process. It is then this process that counts as a description of the real cognitive process given that it is similar to the cognitive process in some relevant respect. This makes sense for dynamicists who are interested in how things change in the first place and have little interest in those states that are the medium of change (*ibid.*, p. 621). However, as Giunti notes (1995), the instantiation relation insures, at most, a certain *similarity* between the changes that the model aims at describing and what counts as a description of them. Accordingly, DST explanations depend on models that are instantiated in real systems in a *weak* sense: the instantiation/simulation relation insures only similarity between a certain cognitive process and the corresponding simulating process. Setting up correspondences on the basis of predictions between numerical sequences contained in the model and those of the real system's data is not sufficient to successfully explain the data. Rather, mathematical variables need to be identified in the physical substrate of the system for us to say that they have real counterparts in the system's performance. To do this, we need to study the material substrate that implements the cognitive process.

A further reason why DST advocates treat mathematical formal models as adequate explanations of cognitive behaviour has to do with the fact that these models have already been proven explanatorily successful in other branches of science. I have claimed above that an important goal of DST is that of unifying phenomena under a same description and that the achievement of this goal is considered an important step to naturalise cognition. Nevertheless, one can agree that an important pay-off of dynamical modelling is to reveal the existence of widespread patterns (e.g. the core equation of the HKB can be used to describe similar coordination patterns in other physical systems) and, at the same time, deny that the mere fact that dynamical descriptions apply to various physical systems bears on whether they explain the phenomena in question or not:

> "If we want to know why humans exhibit the phenomenon described in the HKB model, it is merely suggestive to note that a similar pattern is observed in a variety of other systems. Given the pattern alone we have no better idea than we had as to how it is that humans (or any other system for that matter) behave in compliance with the model. If anything, then, the broad scope of certain dynamical models merely indicates that many other similar phenomena require explanations as well, and perhaps these explanations will be similar." (Kaplan & Bechtel, 2011, p. 441)

A naturalistic account of cognition, I will argue, doesn't necessarily have to dispense with notions that, *prima facie*, don't belong to the natural world (e.g. the notion of representation). As I start arguing in the rest of the chapter and more in detail throughout the thesis, there are cases where appealing to representation is inevitable to make cognitive behaviour intelligible.

2.3.2 – Rethinking the format of representations

According to DST, a cognitive behaviour can be properly explained in terms of lumped parameters and differential equations without the need to posit representations. In particular, the abandonment of the notion of representation is intended to facilitate the study of complex phenomena that unfold over time.

In this section I argue that a complex behaviour that unfolds over time can be explained in representational terms too. This undermines the DST claim that the abandonment of representations is necessary to account for dynamics in the study of cognitive phenomena. On the contrary, both anti-representational and various representational approaches share the idea that timing and dynamics are essential in cognition.

Consider connectionist approaches. They constitute an important category that anti-representationalists fail to adequately address: they are deeply representational although they do not employ the traditional classicist notion of representation according to which representations are "quasi-linguistic" structures whose contents are strings of symbols operated on by a read/write/copy architecture (Churchland, 1989; Clark & Toribo, 1994). Rather, connectionists consider internal representations to be vectors of activations in neural networks, whose processes are vectors' transformations in high dimensional state spaces. As Andy Clark and Josefa Toribo clearly show (1994), the main disagreement between classicists (e.g. Smolensky, 1988; Fodor & Pylyshyn, 1988) and connectionists lies in the identified format of internal representations rather than in the existence of representations in the first place.

Connectionist representations are less transparent and sequentially manipulable than classicist representations. While quasi-linguistics forms of representation operated upon sequential processes seem inadequate to account for dynamics and couplings (i.e. Clark, 1993), the same cannot be said for connectionist forms of representation. Indeed, there are "kinds of fast, efficient coupling often achieved by connectionists neural network style solutions […]; solutions which are nonetheless recognised as falling into a more generally representationalist camp" (Clark & Toribo, 1994, p. 412). Paul Churchland, for instance, treats connectionist networks as representational systems embodying knowledge structures in the form of *prototypes*. Prototypes are points or small volumes that can be depicted in abstract state spaces of possible activation vectors and that can be given a dynamical description in terms of "attractors".

Rather than attacking the notion of representation as such, then, the DST seems to criticise a specific type of explicit and linguistic-form representation. If this is the case, then, there is room for genuinely representational systems that employ non-sophisticated forms of representation to perform their tasks successfully. In the next chapter (see also chapters 6 and 7) I will examine examples where this more plausible reading of representation, together with its explanatory pay-offs, becomes clearer. At present, the current discussion suffices to say that embracing a dynamical perspective doesn't force us to give up on representationalism; rather, it opens up the possibility to rethink and enrich the notion of representation, in particular with respect to its format and its role in the cognitive economy.

DST also forces us to examine which tasks do indeed employ representations and which don't: not all behaviours need explanations that posit representations. The behaviour of the Watt Governor, for instance, can be explained by employing a purely causal story about its workings. Nevertheless:

> "The success of a non-representational analysis of a device like the Watt governor […] fails to argue for a more generic anti-representationalism. For since the dimensions of the relevant state space were straightforwardly physical (available without significant computational effort from the ambient environmental input), the result is effectively trivial. By contrast, as soon as we are dealing with state spaces whose dimensions are more abstract, and hence cover a superficially very disparate range of patterns of physical stimulation (as in e.g., responding to an item as "valuable", or even detecting the presence of a given phoneme (see Seidenberg & McClelland, 1989)), the dynamical system story becomes a representational one (too). Thus, unless you believe that human cognition somehow operates without recording gross sensory inputs so as to draw out the more abstract features to which we selectively respond, you will already be committed to a story in which the states spaces themselves are properly seen in representational terms." (Clark & Toribo, 1994, p. 423)

In the last section of the chapter I will introduce examples from the so-called "representation-hungry" problem domain to show how the behavioural examples that DST considers might not require a representational gloss, while others do.

To summarise what I have claimed so far, my arguments against the explanatory goodness of DST explanations are the followings:

- Formal mathematical models are useful for making predictions about the unfolding of a system's behaviour over time, but insufficient to explain it. I claimed that a genuine explanation of cognitive behaviour requires the identification of its responsible mechanism too. This, in turn, requires a deeper analysis of the implementational substrate that realises the cognitive behaviour
- Formal mathematical methods can be employed to unify disparate physical phenomena under common generalisations, but they are not sufficient to explain the phenomenon in question
- DST instantiation relationships between formal models and physical systems are based on models' predictive powers alone, hence they are insufficient to explain the real processes responsible for behaviour
- Naturalising cognition doesn't require the abandonment of the notion of representation; rather, complex cognitive phenomena can be adequately accounted for by employing a non-classicist notion of representation.

2.4 – Behavioural Systems Approach

Fred Keijzer (e.g. 1998, 2005) takes a different route to show that cognition and cognitive behaviour can be properly explained without positing internal representations.

He argues that cognition needs to be reconnected to the world, and that, to do this, a necessary first step consists in rejecting the idea that cognitive systems employ representations as stands-in for things in the environment. Indeed, a major tenet of anti-representational accounts is the claim that many cognitive processes are closely and intrinsically dependent on external ones and on the dynamical interplay between internal processes and bodily and environmental characteristics. Work in robotics (e.g. Brooks, 1991) is often used to explicate this claim. Robots function properly when there is a specific control structure, a specific embodiment and a specific environment. The internal control structure itself is not sufficient to make sense of the workings of the robot; rather, the body and the environment are essentials to generate intelligent behaviour.

Within Keijzer's Behavioural Systems approach, a behavioural system is understood as a neural, bodily and environmental interaction system. In contrast to behaviourism that points out only the *functional* regularities of behaviour[15], Behavioural Systems explanations of cognitive phenomena also require the identification of internal (structural) mechanisms. In particular, from a Behavioural Systems perspective, a good structural characterisation (i.e. a structural-anatomical description of the physical behaviour itself) is necessary to uncover the mechanisms underlying cognitive behaviour:

[15] "In psychology, the behaviorist solution of behavioral description can be described as function without structure. Structure has been shifted out as irrelevant for operant behavior. For example, changes in the feeding behavior in insects and rats can be functionally equivalent while the structural properties of the behaviors are hugely different. […] Such functional aspects remain the dominant way of characterizing behavior. Movements tend to be treated as 'motor behavior', another functional form of behavior. In contrast, there is little elaboration of the idea that movements—or rather sensorimotor couplings in the embodied view—ought to be taken as the general and basic structural components of behavior's functional regularities." (Keijzer, 2005, p. 131)

> “[…] functionality emerges from the system, and is not a basic feature. A structural characterisation stresses (neural, bodily and environmental) subsystems and the generation of behavior.” (Keijzer, 2005, p. 132)

Consider an insect’s wing beating behaviour. According to Keijzer, insect wing beating is a cognitive behaviour that can be explained by appealing to on-line and off-line components: it is controlled by both on-line external feedback from the wing’s beating movements and by off-line internal oscillators that set a basic rhythm independently from such feedback. The on-line components are crucial to the insect’s behavioural success and correspond to the couplings between the insect and its environment; the off-line components refer, instead, to certain processes internal to the insect that are simultaneous with the on-line ones.

The centrality of the on-line components and the importance attributed to the dynamics of the system are the main building blocks of the Behavioural Systems approach, as clarified in the following quote:

> “Cognition to a large extent depends on, or, some would hold, even consists of perception-action loops that build organism and environment together in a continuous reciprocal interaction.” (*ibid.*, p. 124)

Cognitive processes depend on external factors and on the dynamics governing the complex interactions between the system and its environment so deeply that, according to Keijzer, we can say that cognition *consists* of perception-action loops and dynamics.

Here again, we recognise the peculiarity of this approach with respect to the Classical Computational Theory of Mind: the Behavioural Systems approach attributes a central role to bi-directional, circular, loopy structures of sensorimotor and perception-action couplings in the explanation of cognitive behaviour. The cognitive behaviour itself is not seen as the result of an internal process, but as an intermediate step in an ongoing series of perception-action loops. There is therefore no need to posit internal representations: off-line components are internal states with no representational content.[16] The wing-beating behaviour can then be explained as a joint set of adaptations (i.e. the insect wing beating gets modified according to the external circumstances) and dynamics (i.e. the behaviour of the insect unfolds in time and consists of a series of events that occur over time).

Since the notion of internal states applies widely — Keijzer speaks of "universal presence of neural and other regulatory factors that modulate all ongoing behavioural processes" (*ibid.*, p. 139) — on-line and off-line processes characterise any low-level as well as high-level cognitive behaviour. Consider the following quote:

> "Given these considerations it makes no sense to cast perception-action coupling as on-line and cognitive processing as off-line […]. And, it would be a case of representational overstretch if the traditional notion of representation is applied to all kinds of internal states required to account for different aspects of the whole spectrum of behaviour, from bacterial behaviour to the most complex cognitive tasks." (*ibid.*)

The behaviour is cognitive *simply* because: (i) its outcome is adaptive; (ii) it results from the dynamic interplay between the organism and the environment.

Showing that cognition results from processes and components which are not "special" compared to any other natural process is an important goal for Keijzer and, more in general, for anti-representational approaches. Achieving this goal would imply that there exists a framework to study and explain cognition as a purely natural phenomenon. Keijzer is confident that such a naturalistic account of cognition is

[16] It is interesting to note here, as I will clarify later on in the chapter, that the role of these off-line processes operating on internal non-representational states is different from other roles typically attributed to off-line processes in the representational literature. The latter usually characterises as off-line those processes that stand-in for something that is outside in the environment or for something that is currently absent. This more common usage of the notion of off-line processes is clearly related to a representational function of such processes. For more details on this second reading of off-line, see Rick Grush's emulation theory of representation (2004).

possible and that the rejection of the notion of representation is a necessary step towards it. Internal states do not stand-in for aspects of the environment. They are essentially enabling and regulatory components of cognitive processes:

> "Rejecting or criticizing the use of representations is then taken to imply a view that solely relies on the immediately present environment for guiding perception-action couplings. However, within a behavioural systems approach one can, and must, acknowledge the need to incorporate internal states as relevant factors in the off-line guidance of perception-action couplings. […] Acknowledging internal states does not require a commitment to a representational interpretation of these internal states. […] At first sight, one may tend to equate the behaviourally relevant internal states of a cognitive system with something like intentional states or cognitive representations." (Keijzer, 2005, pp. 138–139)

How can a Behavioural Systems approach deal with classic cognitive topics, such as playing chess (*ibid.*)? Keijzer's answer consists in showing that rejecting representations doesn't imply valuing only the on-line components of cognitive behaviour. Rather, appealing to internal states is necessary to explain on-line behaviour.

2.5 – Criticisms of Behavioural System Approach

To sum up, Keijzer uses the following arguments to argue against representations:

i. A cognitive behaviour consists of on-line and off-line processes (i.e. neural and other regulatory factors)

ii. Explanatory power is achieved once it is possible to highlight the universal features that cognitive processes, and consequently cognitive systems, share with other natural processes and natural systems

iii. Cognition depends *to a large extent* on perception-action loops, which are interpreted as on-line components

Let me consider each of these points individually.

The idea that cognitive behaviours depend on both on-line and off-line processes (i) is widespread even among representationalists, who usually argue that cognitive systems need to rely on internal processes to support and complement on-line processes. This is particularly evident when the explanandum phenomenon doesn't depend closely or directly on available environmental states (e.g. imagery, memory, planning or reasoning). In these cases, priority is given to internal features rather than to perception-action loops: advocates of representationalism say that systems can behave cognitively even in the presence of reduced or absent relevant information from the environment because they employ internal representations. An account that cited only on-line and off-line processes over regulatory states wouldn't be sufficiently explanatory. The role played by off-line components and processes in representational accounts is different from the role they play in a Behavioral Systems account: off-line components are not only used by the system to regulate on-line dynamics, but function as surrogates for certain on-line features. Consider Rick Grush's account of mental imagery (2004). Mental imagery is a cognitive behaviour whose explanation requires considerable attention to its off-line components and processes. How could this behaviour be explained within Keijzer's account? Could we say that it is the result of perception-action loops? Could we treat its internal states as purely regulatory neural states? A representational story seems to be more explanatory here: the agent can imagine an event without performing any overt behaviour because the underlying process, which is not a perception-action loop in Keijzer's sense, is based on the internal manipulation of states that function as surrogates or stands-in for some (absent) environmental features (see also Jeannerod, 1995). As I will briefly argue in the last section of the chapter and more in the next chapter, behaviours that require off-line representational components are more common than we might imagine at first.

Following the same kind of reasoning and considering now point (ii), I believe that in order to identify what is typical of cognitive behaviour we cannot only highlight features that are common among all physical systems interacting with the environment. If we want to explain decision-making processes, for instance, we want an explanation that doesn't only show what is in common between a decision process and, for example, a motor process; rather, we want an explanation that can show how a system can perform decision processes in the first place and what distinguish these processes from others. In other words, we do not have a good explanation of a cognitive phenomenon when we show how it is similar to other physical phenomena, but when we can identify its responsible mechanism — see my criticism of DST in 2.3.1.

Let me now consider point (iii) and the idea that cognition depends *to a large extent* on perception-action loops.

This is another claim that I think misses the mark, as almost every representationalist would agree with it. Indeed, many advocates of representationalism (e.g. Bechtel, 1998; Clark, 1997; Grush, 2003) understand perception and action as cognitive. Extending the cognitive domain so as to include also lower-level behaviours is surely an important pay-off of anti-representational and dynamical approaches, as I argued above with respect to DST, but not a claim that can be used to argue against representationalism in general. Indeed, Keijzer himself recognises that the importance of loopy structures is not confined only to anti-representational accounts of cognitive abilities, but widely adopted in cognitive science (Keijzer, 2005, p. 134). On the one hand, focusing on dynamics and loops does not force us to reject the notion of representation as explanatorily useful and, on the other hand, the importance attributed to perceptual and motor processes and to system-environment interactions does not require a commitment to the thesis according to which every behaviour is representational. Considering again the wing-beating behaviour discussed above, saying that it depends on perception-action loops and on insect-environment dynamics doesn't force us to conclude that it also results from processes operating on internal representations. Showing that certain behaviours, which result from dynamics and perception-action loops, can be explained in non-representational terms is not sufficient to claim that every cognitive process doesn't require representations to be adequately explained.

2.6 – Representation-hungry problems

One substantial trouble affecting DST and Behavioural System approach is that the kinds of problem-domain invoked are just not sufficiently "representation-hungry" (Clark & Toribo, 1994, p. 418). Rather, they are, without exception, domains where environmental stimuli can be used in place of internal representations. It is therefore unfair to use these cases to illustrate a more general anti-representational claim.

Clark and Toribo name a "representation-hungry" problem domain any domain where one or both of the following conditions apply:

- The problem involves reasoning about things or states of affairs that are not directly present, non-existent or counterfactuals
- The problem needs the system to be sensitive to environmental parameters that are complex or ambiguous

The ability to track the distal or the non-existent, for instance, requires, *prima facie*, the use of some inner resource that can be employed to allow appropriate behavioural coordination without the constant guide of inputs from the environment. Domains where this ability is required are less rare than we might expect. Reasoning about absent environmental features or counterfactual scenarios is central in many behaviours. Non-language animals, for instance, seem to anticipate the movements of pursued preys and to engage in counterfactual reasoning when selecting how to grasp food.

Behavioural success in animals, including humans, often depends also on the ability to compress or dilate input space. In some cases, animals need to interpret inputs that are quite similar as deeply different and inputs that are quite different on their immediate coding as deeply similar. This suggests that animals are able to isolate only that information contained in the inputs that is relevant to their coordination with the current environment. The internal states developed to serve this end are, according to Clark and Toribo, just internal representations whose contents concern the states of affairs thus isolated.

Even basic visual abilities (e.g. object recognition) may require the use of similar strategies, as recent neuroscientific research has shown. The ability to recognise the object from any one of a number of distances, angles, settings, and so on is best explained by supposing that the system first transforms the input into a canonical representation frame and only then matches this transformed product to its stored knowledge to carry out the identification task.

All these cases are, interestingly, more widespread than we might have originally supposed. Given that they all require the employment of representations, it is hard to agree with the general anti-representational framework I have considered thus far.

2.7 – Conclusion

In this chapter I analysed the anti-representational framework by examining two anti-representational accounts: Dynamical Systems Theory and Behavioural Systems approach. I then offered various reasons why the anti-representational framework cannot provide good explanations of cognitive phenomena.

According to the authors I considered, a good explanation of cognitive behaviour is a naturalistic explanation that shows the place of cognition in the natural world. I argued that, despite being an important pay-off of anti-representational explanations, this goal does not free us from the need to give an analysis of the mechanisms underlying cognitive behaviour. In particular, I showed that predictability and unification are not sufficient criteria for good explanations, and I argued that they need to be complemented with the identification of mechanisms. Such additional component allows to distinguish descriptions from explanations and to identify a clearer bridge between models and modelled systems.

In addition to this, I claimed that a naturalistic account of cognitive behaviour doesn't need to reject the notion of representation. Rather, such notion appears to be inevitable to explain a wide range of cognitive phenomena that belong to the so-called "representation-hungry" domain and that do not result from direct couplings with the environment.

Chapter 3 - William Ramsey and the Partial Eliminativist Account of Representation

3.1 – Introduction

In the previous chapter, I examined two anti-representational proposals and their arguments against the usefulness of the notion of mental representation in explaining cognition. In the current chapter I will discuss a different attack on the notion of representation put forward by William Ramsey (2007).

According to Ramsey, a genuine representational account explains the cognitive success of a system in terms of internal representations and operations over them. For a state to be a representation, he says, it needs to satisfy certain desiderata. The result of his analysis is a partial eliminativist thesis according to which only the Classical Computational Theory of Cognition (CCTC), but not the newer accounts (i.e. connectionism and cognitive neuroscience), is genuinely representational. The structure of the chapter is as follows.

I will first define what Ramsey calls the *job description challenge* that sets the standards that a theory needs to meet to be representational. According to Ramsey, the notion of representation employed in the CCTC can meet this challenge, while the so-called *receptor notion* employed in connectionism and in cognitive neuroscience cannot. Receptor states are reliable causal relays rather than representations.

I will resist Ramsey's partial eliminativist thesis by drawing on various arguments. I will first claim that the central distinction that Ramsey correctly highlights between theories of representation and theories of content does not apply in the case of the CCTC. In particular, I will show that the CCTC is representational because it explains the cognitive success of a system in terms of internal models that the system can employ to draw inferences about the world (i.e. it is model-based), and not because it adopts an isomorphism-based theory of content. In addition to this, I will argue that the isomorphism-based theory of content is inadequate.

I will then claim that connectionist and cognitive neuroscientific explanations are genuinely representational because they often explain the success of a cognitive system in terms of the exploitation of an internal model, whose representational components and relations allow the system to reason about the world.

3.2 – The job description challenge

The first step in Ramsey's argument consists in identifying a list of minimal criteria for something to be a representation; the second step involves the use of this list as a benchmark to judge whether cognitive theories are justified in talking about representation. Ramsey argues that if a theory employs a notion of representation that doesn't match these criteria, then that theory is not representational. Connectionist and cognitive neuroscientific accounts are not representational because they employ a notion of representation — the so-called receptor notion of representation — that doesn't meet these criteria and that doesn't yield any explanatory benefits over and above that of a reliable causal relay.

Folk psychology provides the list of minimal criteria for something to be a representation. As I have already described in chapter 1, the folk-psychological framework of explanation characterises human cognition in terms of propositional attitudes (e.g. beliefs, desires and intentions, whose content can typically be expressed in propositions) and their causal relations.

Within folk psychology, the notion of representation is defined not only in terms of its content, but also in terms of the functional role it plays within the system: for a state to be a representation, the state should have content and it should use that content in a way that is consistent with deductive principles. Accordingly, Ramsey identifies the following minimal criteria for something to be a genuine representation:

1. A representation has non-derived intentional content
2. A representation plays a causal role

To these two features, Ramsey adds a third one:

3. The causal role that a representation plays is dependent on its intentional content

Ramsey's main thesis is that only a theory that invokes representations whose features match these minimal criteria can be considered representational in a genuine sense. This is called the *job description challenge*.

A theory can meet this challenge when it offers reasons for why certain internal elements are *genuine* internal stands-in for external features. This means that a theory that only accounts for how certain internal states gain their content or a theory that only accounts for how internal states function as stands-in is not a genuinely representational theory.[17] Consider a compass. Ramsey claims that a compass is a non-mental representational object because the position of its needle informs cognitive agents about directions (i.e. the position stands-in for possible directions).

[17] Nevertheless, Ramsey seems to believe that a theory that can show how certain internal states function as representations is more likely to be genuinely representational than a theory that can only provide a story about how internal states gain their content.

The position of the needle entails facts about the world because it is nomically dependent on magnetic north (i.e. it acquires its content through a certain nomic dependency). To explain why a compass functions as a representational device, Ramsey argues, we need both stories. Simply knowing how the needle acquires its content wouldn't suffice: a person might understand the needle's nomic dependency on magnetic north without knowing how the compass actually functions as a representational object. At the same time, merely knowing that the needle's position informs about direction is not enough to be able to use the compass as a representational device.

3.3 – The Classical Computational Theory of Cognition

Ramsey believes that there is only one theory in cognitive science that meets the job description challenge, that is, a theory that can offer not only a theory of content but a genuine full-blown theory of representation. This is the Classical Computational Theory of Cognition (CCTC). In what follows, I will examine how the CCTC is committed to representations.

3.3.1 – IO-representations

The first kind of CCTC representational commitment depends on the compositional nature of computations.

A computational task is generally understood as a sum of smaller computational sub-tasks, each one characterised by its own inputs, operations and outputs. Consider a computational process that transforms numbers into products (multiplication). Ramsey says that:

> "Although we say various mechanical devices do multiplication, the transformation of numbers into products is something that, strictly

> speaking, no physical system could ever do. Numbers and products are abstract entities, and physical systems can't perform operations on abstract entities." (Ramsey, 2007, p. 68)

Given that numbers and products are abstract entities, they can be used by a system only if they are treated as representations of numbers and products: a system can multiply numbers only if it has something internal that stands-in for those numbers. To make sense of this computational process we then need to posit the existence of representations of numbers and representations of products, respectively in terms of inputs and outputs. A cognitive theory that wants to explain how multiplication happens in the brain should therefore account for how the brain can transform representational inputs into representational outputs.

Consider another example. If we want to explain how a cognitive system recognises faces, we need to treat the inputs to the system not as actual faces, but as "some sort of visual or perhaps tactile representation presented by the sensory system. The outputs are also representations – perhaps something like the recognition 'That's so-and-so', or perhaps a representation of the person's name" (*ibid.*, p. 69). Generally speaking, Ramsey believes that cognitive theories should aim at explaining cognitive processes not in terms of physiological changes between events, but in terms of how certain events, which represent for instance faces, get transformed to allow the system to perform successfully in its task.

Within the CCTC, the brain is understood as an information processor: we are "justified in treating a cognitive system's inputs and outputs as representations because, given what we know about cognitive systems, we are justified in characterizing many of their operations as having certain types of starts and finishes; namely starts and finishes that stand for other things" (*ibid.*, p. 70).

In addition to the assignment of representational status to inputs and outputs of the overall process, the CCTC also considers the inputs and outputs of its sub-processes to be representational:

> "If there is an inner sub-system that is an adder, then its inputs must be representations of numbers and its outputs representations of sums. If these internal structures are not serving as representations in this way, then the sort of task-decompositional analysis provided by the CCTC doesn't work." (*ibid.*, p. 72)

Within the CCTC, Ramsey notes, treating sub-computations as representational is necessary to explain why cognitive systems can do multiplication, recognise faces, and so on. As these representations are internal to the system and characterise the inputs and outputs of sub-computations, Ramsey calls them IO-representations. He then points out that their content is essential for the causal role they play in the cognitive system: they need to be about the relevant computational arguments or values to allow the sub-computations to do their job in the overall computational process. Although these representations "don't accord with our commonsense understanding of *mental* representations, they nevertheless play a functional role that is intuitively representational in nature. Their role is intuitively representational because we recognise that the systems doing addition, or comparing chess moves, treat their inputs and outputs as symbols standing in for things like numbers or chess game scenarios. […] the CCTC invokes a notion of internal representation that […] is actually built into the fabric of its explanatory framework and thereby does essential explanatory work" (*ibid.*, p. 74).

3.3.2 – S-representations

The second CCTC representational commitment is via S-representations or Structural-representations. These representations are the components of internal models that cognitive systems employ to successfully perform in cognitive tasks.

Ramsey calls these components "structural" representations because they stand-in for structural features of the target domain by mirroring or by being *isomorphic* to them. The isomorphism here doesn't characterise the relationships between specific internal states and specific aspects of the target domain. Rather, it is the overall internal model that is isomorphic to the states of affairs and that, as a whole, represents. On this point Ramsey argues that:

> "A map illustrates this kind of representation. The individual features on a map stand for parts of the landscape not by resembling the things they stand for, but by participating in a model that has a broader structural symmetry with the environment the map describes. […] some sort of structural or organisational isomorphism between two systems can give rise to a type of representational relation, whereby one system can be exploited to draw conclusions about the other system." (*ibid.*, p. 78)

A map represents certain structural features of the landscape that a cognitive system can use to draw inferences about the world (e.g. finding out the distance between two places). When I visit a city for the first time and I need to get from the station to the hotel, I usually read the map of the city to find out the distance between the station and the hotel and the best direction to take. The map helps me to get to my hotel successfully because it appropriately resembles (i.e. represents) the most crucial structural features of the city, thus allowing me to draw conclusions about the target system (i.e. the city and the location of the hotel).

Ramsey argues that, in the same way in which a cognitive system can use the map of a city to draw inferences about the city, a cognitive system can perform successfully by relying on an internal model of her environment. Isomorphism applies both to concrete external models, such as the map, and to internal models.

Internal models are important features of CCTC style explanations because they allow a kind of *surrogative model-based reasoning*: a cognitive system can successfully draw inferences about the structure of the world by reasoning about the structure of her internal model of the world. If I am in my hometown and I need to buy some bread, I can successfully satisfy my goal by relying on an internal map, whose components and relations mirror (are isomorphic to) those of my hometown.

Mindless Bob

Consider Bob, a cognitive agent who is trying to discover whether and how members of a numerous family are related. He knows the family, but finds it difficult to remember all their kinship relations.

One way in which Bob could succeed in the task is by drawing the family tree on a paper and then connecting the different names with lines. This strategy would allow him to draw inferences about the various kinship relations on the basis of a concrete and visible diagram of the whole family, whose elements would represent the family's members and relations.

Bob could also succeed in the task by employing a different strategy: rather than drawing a diagram, he could form "if-then" propositions. These propositions would stand-in for the relevant elements of the family tree and their connections (e.g. "if X is Y's sister, then X is Y's daughter aunt") and they would allow Bob to find out all the various kinship relations. In this case, we would explain Bob's success in terms of the exploitation of an internal model of the family, whose components and relations mirrored, or were isomorphic to, those of the family. These components would be genuinely representational because they would stand-in for the elements of the target domain.

According to Ramsey, in the same way in which we do not doubt the representational status of names and connecting lines in the diagram, we should not doubt the representational status of the elements of Bob's internal model given that their content is essential for the causal role they play in the cognitive processing (e.g. they need to stand-in for the relevant faces and relations of the target domain). At this point, Ramsey believes we could ask two questions.

We could ask whether Bob succeeds because he relies on an internal model and because he knows the meaning of its elements. Ramsey affirms that the answer to this question would be negative: Bob performs appropriately because he can follow the structure of the symbols and operate on them in a purely mechanical way. There would then be no explanatory gain in calling the elements of the internal model representations. A syntactic story about *how* the symbols get compared and about which rules get applied on them would be sufficiently explanatory.

We could instead wonder *why* those specific symbols and their use make Bob perform successfully. Ramsey says that a syntactic story here would not provide a sufficiently explanatory answer. Rather, we need to understand the symbols as representational elements that stand-in for features of the diagram, and the operations among them as instantiating a sort of surrogative reasoning.

Bob's success in the task can then be explained in a syntactic and purely mechanical way if we want to understand how the process works, and in a representational way if we want to understand why Bob manages to perform successfully by exploiting those processes. In Ramsey's own words:

> "[…] we can't fully understand how mindless-Bob performs the operation of figuring out how two people are related unless we understand his operations as involving the implementation of a model. And to understand his operations as an implementation of a model, we need to look at the elements of these operations – in particular the marks on the page – as representations of people and kinship relations." (*ibid.*, p. 85)

Ramsey's point here is that Bob's internal model is representational although Bob doesn't understand the meaning of the symbols (i.e. that the letters stand for family's members and that their connections stand for their relations). Having the model in place, Bob can succeed in his cognitive task by mechanically following certain rules. He claims, though, that it is necessary that the content of representation gets fixed by isomorphism: representations need to stand-in for something else for the model to work, and they can be stands-in for external features if they share structural similarities with them.

Accordingly, a notion of representation can be explanatory useful even in a purely mechanical problem-solving system and even when it is part of a framework that is not committed to any kind of implementational characterisation of the states and processes that traffic in representations. Implementational details, Ramsey says, might be important to actually construct a machine able to perform computational processes, but not to understand the sense in which a cognitive process is computational.

3.3.3 – CCTC and the job description challenge

Ramsey argues that the CCTC is committed to representations and that the notion of representation it employs meets the job description challenge; hence the CCTC vindicates our folk-psychological understanding of representation. Let me briefly summarise the arguments he addresses to show this.

As described above, Ramsey points out two ways in which the CCTC is committed to representations.

The first commitment is via IO-representations:

- They derive from the compositional nature of computational processes, which are typically considered in terms of their sub-parts, each one understood as a sub-computation
- The content of IO-representations is essential for the causal role they play in the computational processing because they need to stand-in for the relevant computational arguments or values

IO-representations are then part of the CCTC explanatory framework although they are not similar to our folk conception of representation.

The second representational commitment is via S-representations:

- A cognitive system performs successfully in a cognitive task by relying on an internal model of its target domain, whose components are isomorphic stands-in for those of the target domain

- By exploiting an internal model, an agent can perform a sort of model-based surrogative reasoning
- A cognitive agent performs appropriately by exploiting the structure of the internal models' components in a purely mechanical way. Nevertheless, in order to explain why those specific components and the operations upon them enable the agent to perform successfully, we need to understand the components as representations of elements of the target domain, and the operations as instantiations of surrogative reasoning

Once more, because CCTC explanations are model-based, the notion of S-representation is a natural element of this theoretical framework.

We can now ask whether the CCTC vindicates our folk-psychological notion of representation. If this were the case, then the CCTC would not be committed to representations in general, but to representations whose features match the minimal criteria (i.e. for something to be a representation it should have non-derived intentional content and it should play a causal role similar to that of beliefs, desires and intentions in folk psychology).

Ramsey claims that IO- and S-representations do share many features of our folk notion of representation: they both have the kind of intentionality that we attribute to thoughts and the discreteness that folk psychology assigns to beliefs and desires. Consider Sherlock Holmes' reasoning process employed to find out how a victim died (Fodor, 1987). Holmes uses a folk-psychological reconstruction of his thoughts, observations and beliefs. His reasoning process has an argument-form, with premises that yield certain conclusions. Ramsey claims that, since the CCTC is good at explaining these kinds of arguments in computational terms by relying on IO-representations, the CCTC can be seen as a scientific framework that can vindicate folk psychology.

This would be possible also on a different reading of the process, this time involving S-representations: Holmes finds out how the victim died by exploiting some kind of internal model, whose components and relations mirror the events that yield to the victim's demise. Here the solution is found by mentally reconstructing the setting of the crime, its relevant events and their connections. Accordingly, Ramsey says, folk-psychological explanations can be cashed out in terms of models and surrogative reasoning and "If this is correct, then folk notions of mental representations may well be very close to the notion of S-representation proposed by the CCTC. […] While it is hard to see how beliefs could turn out to be mere syntactic states with an unspecified representational role (as suggested by the Standard Interpretation), it *does* seem they could turn out to be representational components of models that our brains use to find our way in the world" (*ibid.*, p. 116).

3.4 – The receptor notion of representation

Ramsey's central claim is that the CCTC is the only genuine representational theory and that the newer accounts (i.e. connectionism and cognitive neuroscience) do not employ any genuine notion of representation: their so-called *receptor notion* can be considered simply in terms of reliable causal relay.

In what follows, I will first characterise the receptor notion and I will then consider two separate objections to it.

3.4.1 – Nomic dependency

Connectionism and cognitive neuroscience apply the notion of representation to states that reliably get activated and co-vary with some external features of the environment.

Accordingly, a state X (internal) is called a receptor representation for Y (external) if the occurrence of X is nomically dependent on the occurrence of Y. Consider this example. Cognitive neuroscientists often say that certain brain cells represent features of the environment when they reliably fire in correspondence to them. This is for instance the case of the so-called "edge cells", that is, neurons in the visual area of the brain that fire whenever agents see edges in the environment. Claiming that some cells reliably fire in response to certain external stimuli is like saying that their activities are nomically dependent on those external stimuli.

Nomic dependency relations are often equated to representational relations in the newer accounts. Ramsey's first critique to the receptor notion lies exactly in this equation: something can be nomically dependent on something else, he claims, without carrying information about it. A cell can fire reliably and co-vary in accordance with certain features of the world without being, at the same time, a representation of those features of the world. Assigning a representational status to this kind of internal states should be avoided because explanatory useless.

Ramsey's first step in attacking the reception-style notion consists in criticising Fred Dretske's naturalistic account of content (1988) as a theory of representation. This move is motivated by the fact that many defenders of the receptor notion of representation agree, at least to a certain degree, with Dretske on what should count as a genuine representational state.

Dretske aimed at offering a purely causal account of what it is for a state to be a representation. His account, however, seemed to leave no room for misrepresentation or falsehood: if X represents Y whenever X reliably co-varies with Y, then it becomes difficult to imagine cases where this nomic dependency does not hold.

To handle this problem, Dretske introduced a teleological component to his account (e.g. Millikan, 1984): an internal state X is a genuine representation of Y not only when it is nomically dependent on Y, but also when it becomes incorporated into the system's processing because of this nomic dependency. States that, for instance, are employed as causes of motor outputs because they indicate, or stand-in for, certain external conditions are genuinely representational states. Here, the informational content of such states is essential to explain why they get incorporated into a system's functional architecture. Misrepresentation can then be explained in the following way: a state misrepresents when it gets recruited in a system for playing a functional role that is different from the one it has been selected for either through learning or natural selection. Coming back to Ramsey's example, if the needle reacted to something different from the magnetic north, the compass would misrepresent, that is, it would respond to something different than what it was recruited for.

The teleological character of representation is for Dretske and his followers not only important for dealing with the problem of misrepresentation, but also for explaining why a certain representation with certain content gets incorporated into a system's processing.

To summarise, Dretske believes that an internal component X is a representation of Y if:

- X is nomically dependent on or reliably co-varies with Y
- X becomes part of the system's processing because it is nomically dependent on Y

Ramsey criticizes Dretske's account on different grounds.

The first critique is against Dretske's claim that the teleological component is sufficiently explanatory for cases of misrepresentation. Ramsey observes that misrepresentation can't be equated with malfunctioning: a device or a state can play a role different from the one it was supposed or designed to play without being a representation. A television, for instance, can malfunction and its malfunctioning can be explained in non-representational terms; the same can be said for many biological processes. We therefore don't have an account of misrepresentation by appealing to teleology alone. Rather, we first have to assume that the state is a representation in order for it to misrepresent, that is, we first need to know in virtue of what that state is representing instead of doing something else. Once we know that we are dealing with a representational state, we can ask how the state has the specific content it has and how it has a different and false content in other occasions. Dretske, however, "limits *what* a state represents by appealing to what it *ought* to represents, leaving the first question about whether the state in question represents *at all* untouched" (*ibid.*, p. 132).

A second problem concerns the relationship between "being nomic dependent on" and "carrying information about" something. Advocates of the receptor notion believe that a state carries information about something when it is nomically dependent on it. But if being an indicator that carries information means for a state to be a reliable responder, then every time we talk about information carrying we could talk about nomic relations between states. Indeed, Ramsey claims that these relations would be such that they could be used by the system to learn about the current states of affairs. If, for instance, we want shade on our back porch at a certain time in the afternoon, we might think of planting a tree at a certain distance from our porch because we know that its shadows will fall exactly where and when we want. In this case, Ramsey affirms, we would exploit the nomic relation that exists between the tree, its shadows and the sun to assign to the tree the job of shading our back porch. It would, however, seem inappropriate to interpret the shadows as representations of the position of the sun in the sky or of the hour of the day because the information that the shadows carry is not relevant to the job they perform: we plant the tree in that position to shade our porch and not to learn about the position of the sun or the hour of the day. We therefore need some additional reasons to conclude that something is employed because of the information that is carried by a certain nomic relation rather than just because it is functional to our purpose. According to Ramsey, one problem with the receptor notion is not that it is explanatory irrelevant in itself, but that the information that results from its law-like relation with external events is explanatory irrelevant.

For all the above reasons, Ramsey claims that for a state X to be nomically dependent on a state Y it does not mean that X carries information about Y. It suffices to say that X functions as a reliable causal relay for Y:

> "A structure can be employed qua nomic-dependent or qua reliable-respondent without being employed qua information-carrier or qua representation." (*ibid.*, p. 138)

The receptor notion of representation is, therefore, explanatory useless because it doesn't provide any additional explanatory pay-off over and above that of reliable causal relay.

3.4.2 – Distributed representational format

Ramsey's second major objection to the receptor notion concerns the causal role that a state should play to count as a representation.

Ramsey claims that having a discrete format is necessary for a state to play the causal role typically attributed to a propositional attitude. One of the problems in the newer accounts is that their representations don't have such a discrete format. The clearest examples can be found in connectionist modelling studies.

Although it employs the notion of computation, connectionism is often considered an alternative to the CCTC because it explains computational processes in terms of distributed rather than localised operations over networks of nodes and connections. Connectionist research has been introduced to better understand how our brains carry out cognitive tasks and, to this end, neural networks have been designed to resemble some aspects of our cerebral neurons and cortical activations. Connectionist networks consist of layers of nodes similar to cerebral neurons. These nodes play different roles and have different values within the overall network. Their values are defined in terms of their "weights": a high weighted connection between nodes has more value than a smaller weighted connection. Neural activations are also characterised on the basis of "thresholds": if the connection's weight between two nodes, A and B, is above a certain threshold, then the activation passes from A to B, and B gets excited. Weights between connections play an important role in the network's economy because it is through their modulation and adjustment that the network learns how to perform a task.

An important feature of connectionist networks is that the information that passes and gets transformed is not identifiable in any specific node; rather, information is distributed across nodes' activations. This means that the semantics of the network depends on relations of similarities and differences among activation patterns through time. If two networks get activated in similar ways, then they encode similar information, and if a same network gets activated in a similar way at time t_1 and t_2, then the network encodes similar information at both times.

Since neural networks compute over information that is encoded in a distributed way, computational processes in connectionism can't be defined in terms of their inner sub-computations.

Importantly, the reason why researchers moved away from the CCTC was to suggest more neurally-constrained explanations of cognitive capacities. CCTC explanations, for instance, are said not to account for the plasticity shown by neural networks. While in artificial neural networks the overall networks can carry out cognitive tasks with only few output problems even in cases of malfunctioning or loss of individual nodes, in the CCTC the disruption of only one symbol has a much more widespread effect on the overall functioning of the system. However, since cognitive systems typically manage to perform cognitive tasks even when their underlying processes are somehow disrupted, connectionist explanations have been considered more appropriate for modelling and explaining real cognitive behaviour. The constraints that the implementation level imposes on connectionist explanations are stringent. A first constraint is dictated by the nature of the brain itself: a very complex net of neurons, connections and activations that vary through time under many different internal and external conditions. To test biologically plausible hypotheses about how the brain manages to carry out complex computational tasks, a first necessary condition is to treat neurons and their connections as the main building blocks of cognitive systems' success. The main motivation behind this shift in format is then based on the connectionist attempt to gain some neural plausibility.[18]

Distributed formats, however, prevent connectionism from meeting the job description challenge: connectionist representations are not causal in the same way in which propositional attitudes are causal in folk psychology.

For cognitive neuroscience, the situation is even more complicated. Although there is no consensus over the format of neural representations (e.g. single cells' activations, grouping of cells, and so on), most neuroscientists and cognitive neuroscientists believe that representations in the brain do not have a discrete format.

Before discussing Ramsey's critiques concerning the receptor notion of representation, let me briefly summarise them:

- A receptor representation is an internal state X that co-varies or that is nomically dependent on an external state Y
- Nomic relations should not be equated to representational relations: something can be nomically dependent on something else without carrying information about it
- A state can be an indicator of another state by depending nomically on it. But, if being an indicator that carries information means for a state to be a reliable responder, then every time we talk about information carrying we could talk about nomic relations between states
- The receptor notion of representation has a distributed format, hence it cannot play the causal role attributed to representation in folk psychology
- Theories that employ the receptor notion do not meet the job description challenge. This means that the representational talk should be avoided and

[18] There is clearly much more to say about connectionism, but for the current discussion we only need to stress the representational format in connectionist networks.

> that the notion of representation should be substituted with the notion of reliable causal relay

In the next section I will examine whether we are justified in concluding, as Ramsey does, that connectionist accounts are not representational. I will argue that some of Ramsey's reasons in favour of CCTC representationalism can be also applied to connectionism and cognitive neuroscience. In the last section of the chapter I will then claim that, while a distinction between a theory of content and a theory of representation is essential, we do not have enough reasons to embrace Ramsey's partial eliminativist thesis.

3.5 – Internal models and representational commitment

Ramsey defends the CCTC as a genuine theory of representation on the basis of two main reasons: CCTC's commitment to IO-representation and CCTC's commitment to internal models and S-representations.

The first commitment has to do with the compositional nature of CCTC explanations: complex computations are explained in terms of simpler sub-computations and, given that each computational process is characterised by non-mental representational inputs and outputs, inner sub-computations need to be characterised in terms of their inner mental inputs and outputs. These inner computational symbols have to stand-in for external features of the environment.

The second commitment is via S-representations. Ramsey shows that the CCTC explains the success of agents in cognitive tasks by positing the presence of internal models, whose elements (S-representations) and causal connections can be exploited to implement surrogative reasoning. This form of reasoning enables systems to draw inferences about the world. However, for these internal models to implement surrogative reasoning, their elements need to be representations of features of the target domain and need to acquire their content by being isomorphic to them. By employing S-representations (and also IO-representations), the CCTC can be considered a genuine theory of representation.

There are two points that I want to raise here. The first point is that S-representations are not peculiar of the CCTC only. Representations are elements of internal models and are isomorphic to elements of the world even in some connectionist networks. The second consideration is that a theory can employ S-representations without embracing the isomorphism theory of content.

3.5.1 – S-representations in connectionism

A commitment to S-representations is not unique to CCTC-style explanations. Connectionism, for instance, employs them too. I consider here only two examples that help clarify this point.

Paul Churchland (1998) studied two identical feed-forward networks trained on the same corpus of 100 photos of each of the 100 members of 4 extended and multigenerational families. The networks were trained so that they could distinguish any input photo as a member of one of the 4 families. After the training period, the two networks were able to generalise successfully to any new example of the 4 families with a degree of accuracy higher than 90%. Churchland explained the behaviour of these two networks in the following way. The networks became able to distinguish each new example as belonging to a specific family because the activation-space of each network became partitioned in a way that, for each of the 25 faces of the first family, there was a specific number of points that tended to be assembled in a given sub-volume within the overall space. This process is often called *clustering*. The same cluster, this time in different sub-volumes of the activation space, characterised the faces of the other 3 families. The clustered and recognisable points of the two networks characteristic of a specific family were called *prototypes* or *concepts*. Prototypical positions encoding information about family members identified the standard causal response to a typical face that belonged, for instance, to family 1. This means that the activation pattern that was said to represent a given member as belonging to a specific family was nomically dependent on (or reliably co-varied with) the member it stood-in for. The notion of representation employed here clearly was the receptor notion.

In discussing the results of the study, Churchland pointed out that the two networks managed to successfully solve the problem although they relied on different coding strategies. The reason for this is that there were similarities in the relative positions of the points (nodes) in the two networks that stood-in for a specific family. What explained how the two networks succeeded in the cognitive task was then some kind of structural feature: the two networks used the degree of similarity or of difference recognised among faces of members of different families to associate a new face to the appropriate family. Indeed, the hidden spaces of the networks got partitioned through training so as to reflect, in a systematic way, the structure of the environment: systematic distance measures stood-in for important family relations.

Following Ramsey's definition of S-representation (i.e. internal states that acquire their content by being isomorphic to their target domain), connectionist explanations can rely on S-representations, hence they can be genuinely representational.

I would like to consider a second example that has been recently offered by Oron Shagrir (2012). Shagrir claims that oculomotor control can be explained by referring to S-representations, internal models and recurrent neural networks: the brain controls the eyes by implementing an internal model in the form of a recurrent neural network with multistable states, one for each eye position. This means that the brain controls the oculomotor system by employing a short-term memory of the current eye position, which is understood in terms of a recurrent neural network with multiple states (S_1, S_2,…, S_n), each representing an eye position (E_1, E_2,…, E_n). The dynamics of this network are such that, whenever the eyes move from position E_1 to position E_2, there is a transition from state S_1 to state S_2 in memory. Certain computational processes operating on the input representing the eye position make these dynamics possible. In particular, for every eye position, there is a corresponding stable pattern of activation in memory. No single cell in memory, or state in the network, represents; rather, it is the collective activation of states that represents a specific eye position: whenever a new stimulus arrives, it perturbs the memory network, thus moving it from the stable pattern of current activated points to a new one. This new stable position represents the current eye position. Interestingly, the distance between two stable patterns of activations in the memory network mirrors that between the two corresponding eye positions. Shagrir claims that the memory functions as a kind of internal map that the system can go and "look up" for solving problems. Accordingly, the explanation of how the oculomotor system controls the eyes refers to the fact that it internally implements and uses a model of the dynamics and positions of the eyes.

In this case it seems that we can conclude that the explanation of how the memory network controls the eyes is genuinely representational: the network performs successfully by employing an internal model, whose elements stand-in for those of the target domain (i.e. the eye positions) and whose connections stand-in for those of the target domain (i.e. the distance between two eye positions). It then follows that these internal representations can be said to be isomorphic to the elements of the target domain. All this is possible, though, in a non-CCTC framework that employs distributed representations.

3.5.2 – The inadequacy of isomorphism-based theory of content

The second consideration concerns isomorphism as a theory of content. As Mark Sprevak (2011) has pointed out, one can embrace S-representations without embracing the isomorphism-based theory of content; the latter is inadequate for representations.

While a representational relation is an asymmetrical relation (i.e. a state can represent another state of affair without that latter representing the former), an isomorphic relation is a symmetrical relation (i.e. a similarity in structure between two objects or two states is symmetrical — if the word "dog" is isomorphic to the word "god", then the word "god" is also isomorphic to the word "dog"). An internal state can then represent an external feature without being isomorphic to it. Isomorphism brings indeterminacy in content: the words "dog" and "god" are isomorphic although they identify very different states of affairs.

If we say that a state is a representation when it is a component of an internal model that the system uses to draw inferences about the world, then we are making a claim about the role that the state plays in the overall cognitive economy of the system. It is a separate question that concerning how the state gains the content it has. The arguments Ramsey addresses in favour of S-representations should then be separated from those used to argue for an isomorphism-based theory of content. Once we separate the two arguments, we are in a position to recognise that many connectionist and cognitive neuroscientific explanations of cognitive performances do employ S-representations, without necessarily embracing also the isomorphism-based theory of content. Consider once more the edge cells example discussed above. In explaining the success of cognitive systems in recognising (i.e. drawing inferences about) edges between light and darkness in their visual field, researchers typically interpret the activity of certain cells in the visual cortex as representing distal edges. By interpreting the activity of these cells in this way, they can account for how a cognitive system manages to perform successfully: these cells represent distal edges and they are part of a wider internal model that the system can "look up" for its environmental success.

A theory can then be representational without embracing isomorphism. As Ramsey correctly points out, it is important to show that a theory is actually committed to representations. Ramsey's critique of Dretske's account of representation consists in showing that a theory of content is not sufficient for a theory of representation: we first need to say why a state is a representation in order to say that it has a certain content and that it can misrepresent. A state can therefore be labelled as a representation rather than as a reliable causal relay if it plays an explanatory role in explaining successful behaviour. In this specific case, if a theory employs representations to talk about the components of an internal model that the cognitive system uses to perform successfully in the environment (i.e. if a theory is committed to S-representations without the isomorphism story), then that theory is representational.

A less ambiguous term than S-representation, which immediately suggests the idea of structural isomorphism, could be of help here. I therefore suggest we could talk about M-representations or Model-representations to identify those states of internal models that a system can employ to draw inferences about the world. Saying that a theory is committed to M-representations means that the theory can explain a cognitive ability in terms of the exploitation of an internal model. The states of this internal model are representations, no matter how they acquire their content.

Accordingly, every theory that posits models as surrogates of (aspects of) the world in reasoning is a representational theory. Connectionism is then committed to genuine representations. If we consider again Churchland's example, the networks are able to perform the task because, through training, they form an internal model of their target domain. In the same way in which Bob's success in the task cannot be explained only by relying on a syntactic story about the assembly and use of internal symbols, we wouldn't be able to explain the success of the two networks without seeing their clustering processes as representational.

Claiming that a theory of content is not sufficient to offer a full-blown theory of representation is then not enough to argue in favour of representational eliminativism in the newer accounts.

The reasons for this conclusion are twofold. First, Ramsey himself does not deny the receptor notion as a theory of content, or the explanatory relevance of nomic dependencies; rather, he denies the receptor notion as a theory of what makes something a representation. Second, a receptor state can count as a representation because of its explanatory role in explaining the success of agents in drawing inferences about the world.

3.6 – Conclusion

In this chapter, I discussed William Ramsey's recent attack to the explanatory usefulness of the notion of representation in connectionism and cognitive neuroscience and I argued for its inadequacy.

I examined Ramsey's defence of the representational status of CCTC explanations (via IO-representations and S-representations) and his partial eliminativist thesis towards the receptor notion of representation employed in the newer accounts. I then resisted Ramsey's partial eliminativist thesis by showing that:

- Connectionism and cognitive neuroscience often employ S-representations because they both typically explain the success of cognitive systems in terms of:
    - Structural isomorphism between the internal structure of the system that gets activated and the target domain
    - Internal models that the system uses to draw inference and reason about the world
- A theory can employ the notion of S-representation without embracing the isomorphism-based theory of content
- The isomorphism-based theory of content is inadequate because isomorphism is a symmetrical relation that entails great indeterminacy in content
- If a theory is committed to representations when it employs internal models (and M-representations) in explaining cognitive abilities, then it is

representational. The newer accounts are, therefore, genuinely representational

- Even if the receptor notion only offers a theory of content, there are still good reasons to consider these internal states as representations

To conclude, I want to stress that, although it is difficult to come up with necessary and sufficient conditions for a state to be a representation, its is often possible to identify cases where a representational talk is applicable and cases where explanations of behaviour cannot avoid appealing to such a notion without losing important explanatory power. I argued that the newer accounts fall into this category although they cannot (yet) offer a full-blown theory of representation.

Chapter 4 - The Personal-Subpersonal Distinction

4.1 – Introduction

In this chapter I analyse the explanatory aims, methodologies and vocabularies of personal and subpersonal explanations of mental phenomena.

In the first part of the chapter I discuss personal-level explanations and the autonomy theorists' position with respect to it. In the rest of the chapter I address and critically discuss two purely subpersonal accounts: Churchland's eliminativism and Bickle's reductionism.

I examine these accounts and show that none of them succeeds in providing appropriate explanations of mental phenomena: purely personal-level and purely subpersonal-level explanations cannot properly account for cognition.

4.2 – The distinction

We believe that we normally can and do explain other people's behaviour in appropriate ways. If I see my friend Sara applying for a job and I want to know why she is doing it, I can explain her behaviour by making reference to her beliefs and desires: Sara is applying for a job because she desires a job and believes that by applying for it she can get it. And if I see my flat-mate wearing a thick coat and a warm scarf and I want to explain why, I can easily say that she is wearing warm clothes because she believes that it is cold outside and she desires to be warm.

In explaining people's behaviour we also often, correctly, assume that their behaviour is not simply triggered by environmental conditions. Whether it is really cold outside or not, if I see my flat-mate wearing a thick coat, I explain her behaviour by saying that she believes that outside it is cold and that she desires to be warm. I then manage to explain her behaviour by making reference to her mental states, and, in particular, to her belief and desire.

The content of a mental state doesn't simply depend on how the world is (whether it is cold or not outside), but also on how the agent takes the world to be (outside it is cold). For this reason, mental states are called intentional: they are about something.

Another important feature of mental states is that they are not physical objects. If they were, they would mirror the real states of affairs out in the world and we know that sometimes they don't: my flat-mate can believe that outside it is cold, but, in fact, it is not. What characterises the nature of a mental state is that its content can be expressed with a proposition: she believes *that outside it is cold.* We explain behaviour by making reference to mental states with propositional content and the same propositional content can be linked to different mental states. The propositional content "that outside it is cold" can, for instance, be paired with a belief (believe that outside it is cold), with a hope (hope that outside it is cold), with a desire (desire that outside it is cold) or with any other mental state. The resulting pair of mental state and propositional content gives rise to the so-called propositional attitude.

To properly explain people's behaviour we also make reference to the connections between mental attitudes. In particular, we can explain behaviour because we know that propositional attitudes are often rationally related to give rise to behaviour. Being rationally related means that propositional attitudes are related to each other in accordance with principles of rationality (e.g. propositional attitudes should be consistent and they should follow the rules of logic or probability theory).

The fact that everyday explanations of people's behaviour make reference to rational principles introduces a normative dimension into the picture. We can explain behaviour because we compare them with how an agent, with that combination of propositional attitudes, ought to have acted:

> "[We] explain intelligent behaviour by interpreting it as the behaviour of rational agents. The principles of rationality regulating the interpretation of rational agents are normative principles rather than descriptive generalizations (principles that describe how people ought to behave, as opposed to description of how they generally do behave)." (Bermudez, 2005, pp. 42–43)

Given the central role of normativity in our everyday explanations of people's behaviour, propositional attitudes are often called *reasons* for actions.

We commonly rely on this kind of explanations in our life and this seems to suggest that we often can properly explain other people's behaviour in an intuitive, easy and unscientific way. To explain why my flat-mate is wearing a warm scarf I don't need technical instruments or knowledge of her internal wirings.

Explanations that, instead, make reference to people's brains, their internal wirings and the mechanical descriptions that can be given of them, are commonly called *subpersonal explanations*.

Daniel Dennett (1969) coins the terms "personal" and "subpersonal" to clarify the distinction between "the explanatory level of people and their sensations and activities" and "the subpersonal-level of brains and events in the nervous system" (*ibid.*, p. 93).

While the personal level makes reference to whole persons qua rational agents (e.g. Hornsby, 2000), the subpersonal level makes reference to parts of persons, their brains, their activities and components. Descending from the level of persons (i.e. the personal level) to a lower level of explanation consists in decomposing persons into parts, mainly brain parts, and in identifying how the organised operations of these parts can bring about personal phenomena. Explanations at this lower level are descriptive rather than normative because they identify the explanandum’s causal history or the “sequences of events that can be subsumed under general causal law” (*ibid.*, p. 38). Personal and subpersonal explanations appeal then to different vocabularies, to different sets of norms and to different principles.

Dennett introduced this distinction to clarify the explanatory domains of different disciplines. The practice, however, has shown that this distinction has created more confusion than clarity.

There are various reasons for this. The main reason, I believe, has to do with the difficulty in integrating, comparing and contrasting knowledge coming from different explanations of the same explanandum phenomenon. Jose Luis Bermudez refers to this problem as the *interface problem* (Bermudez, 2005), that is, the problem of how folk psychology, the main type of personal-level explanation, interfaces with scientific psychology, cognitive science and neuroscience.

According to Bermudez, personal-level explanations are *horizontal explanations* that aim at explaining a particular state or event in terms of distinct and often temporally antecedent states and events. He says:

> “Suppose we ask why the window broke when it did. An horizontal explanation of the window’s breaking might cite the baseball hitting it, together with a generalization about windows tending to break when hit by baseballs travelling at appropriate speeds.” (*ibid.*, p. 32)

These kinds of explanations are not suitable to answer other kinds of questions. What if I want to know why the window broke when the baseball hit it? How can we know why certain generalisations hold? According to Bermudez, *vertical explanations* are suitable to answer these why-questions because “the project of vertical questions can be broadly characterised as explaining the grounds of horizontal explanations" (*ibid.*, p. 33).

In the rest of the chapter I will analyse three different answers to the interface problem. As I will show, they are all answers that, in different ways, consider the interface problem as a non-problem. On the contrary, I will argue that both the personal level and the subpersonal level are required to adequately explain cognitive phenomena.

I will begin by examining the arguments used to vindicate the explanatory independence of the personal level with regard to the subpersonal one.

4.3 – The autonomy of the mental

Advocates of the autonomy of the mental (e.g. Hornsby, 2000; McDowell, 1994; Davidson, 1963) hold that explanations of mental phenomena become intelligible and can be explained only in the context of our human life (McDowell, 1994, p. 204). They want to show that "what is explained at the personal level *cannot* be explained over and again at a lower level" (Hornsby, 2000, p. 8) because "when we abandon the personal level in a very real sense we abandon the subject matter of [persons' mental states] as well" (Dennett, 1969, p. 38).

Jennifer Hornsby (2000) argues that we need to focus on mental states if we want to explain the behaviour of a person *qua rational agent*. A different approach would simply put the person out of the picture:

> "[…] there is no prospect of finding a person intelligible in terms of physical goings on inside her. If one speaks impersonally, one is barred from the sort of normative account that might show a person's doing something to be understandable. [At the same time] there remains a perfectly good question about how it is that persons have the aptitudes

> and capacities that we take for granted when we see mental causation at work. These are capacities to move one's arm when one wants to, to understand another's words and say some of one's own, to see and hear, to recognize faces and expressions […]. How can we do such things? What properties are there of our brains and nervous systems in virtue of which […] we can do them? […] Sub-personal psychology has plenty of tasks in its own. It addresses '*How'*- questions which proceed from empirical ignorance […]." (*ibid.*, pp. 8–9)

Subpersonal explanations are, therefore, important, but they can provide adequate answers only to "how" questions. The personal level is the only level that deals with "why" questions.

Among the reasons why, according to Hornsby, the personal level should be kept distinct from the subpersonal level we find:

- Lower-level explanations don't target people's behaviour
- Subpersonal-level explanations can answer how-questions, while personal-level explanations concern why-questions
- A normative account is needed to make a person's behaviour intelligible and this account is found only at the personal level

John McDowell (e.g. 1994) argues along the same lines by comparing the personal-subpersonal distinction to the constitutive-enabling distinction.

He discusses the personal-level phenomenon of visual experience. Subpersonal explanations of visual experience, he says, make reference to a series of information processes that take place in a subject's visual system. They tell us, for instance, that arrays of intensities and wavelengths are computed by the visual system to yield a sort of image of a part of the environment. Personal explanations, instead, are concerned with what defines the nature of a visual experience as an encounter with an object. McDowell acknowledges that both explanations are needed for a full explanation of visual experience and still maintains that the role of the visual system (i.e. a subpart of a person) is not to inform the person of something.

To make the point clearer, McDowell comments on a famous paper entitled "What the Frog's Eye Tells the Frog's brain" (Lettvin *et al.*, 1959). He claims that "in the metaphor, our parts talk to one another; they do not, at least in general, talk to us" (McDowell, 1994, p. 195). In the frog's case, the frog's visual system doesn't tell the frog anything beyond the fact that there is a bug-like object in a certain position. The role of the frog's visual system is not to process information, but to react to moving objects in the environment:

> "One sub-froggy part of a frog transmits information to another: the frog's eye talks to the frog's brain, not to the frog. […] What tells the frog things is the environment, making things of itself apparent to the frog. […] What the frog's eye does for the frog is to put it in touch with moving specks in its spatial environment." (*ibid.*, pp. 197–198)

According to McDowell, then, visual processes enable a certain kind of encounter with moving objects by putting the frog in contact with its environment. What the visual system doesn't do, instead, is shed light on what is constitutive of visual experience. In other words, subpersonal explanations can only account for the factors that *enable* personal-level phenomena to come about, but they cannot offer *constitutive* explanations for them.

How exactly should we understand the difference between constitutive and enabling conditions?

4.3.1 – Enabling and constitutive conditions

If we look, for instance, at the way in which Hornsby and McDowell treat the distinction, it seems as if the constitutive nature of a certain mental phenomenon can be uncovered solely at the level of folk psychology. Nevertheless, there are cases where enabling information, as autonomy theorists would call them, does make a contribution to our understanding of the constitutive nature of certain personal-level phenomena.

Consider the case of elementary learning developed by Eric Kandel and colleagues (e.g. Hawkins & Kandel, 1984) and discussed by Gold and Stoljar (1999) as an example of purely neuroscientific subpersonal explanation. This theory aims at explaining two kinds of learning — simple learning and associative learning — by making reference to properties of neurons and, in particular, to changes in synaptic strength due to the production of neurotransmitters in sensory neurons. We can understand these forms of learning as personal-level phenomena because they characterise the behaviour of whole organisms and not of sub-parts of them.

The theory employs a basic model that can be adapted to explain a number of different forms of learning, from habituation to sensitisation and classical conditioning. Habituation and sensitisation are two forms of simple learning. Habituation is a form of learning through which an animal gradually ignores a stimulus when it doesn't bring reward or harm, while sensitisation is a form of learning that develops when an animal starts experiencing a harmless stimulus as noxious after the stimulus has been associated with an aversive one. Kandel and colleagues studied these forms of learning in a simple organism, the marine snail *Aplysia californica*.

They found that the Aplysia's innate gill-withdrawal reflex, which followed a neutral tactile stimulus to the tail, could be habituated by reducing the quantity of neurotransmitter released by siphon sensory neurons on the motor neurons.

In the case of sensitisation, instead, they discovered that the responsible process involved an enhancing change of the neurotransmitter released by the sensory neuron on their target cells. Here, while a mild stimulus to the tail produced a mild gill-withdrawal reflex, a shock to the tail activated facilitator interneurons that synapse near the synapse formed by the siphon sensory neurons on the motor neurons. The role and the activity of these interneurons was that of making the siphon sensory neuron release more neurotransmitter after the stimulation. The next time the siphon sensory neuron was mildly stimulated, more neurotransmitter was produced and released on motor neurons and a stronger gill-withdrawal reflex occurred. This process, which is called presynaptic facilitation, consisted in making the siphon sensory neurons more active via the activity of facilitator interneurons.

Kandel and colleagues found that the process of classical conditioning was quite similar to that of sensitisation. In classical conditioning, an unconditioned stimulus (US) and a conditioned stimulus (CS) became paired such that the response that would normally follow US occurred even in the presence of CS alone. They tested Aplysia's behaviour during repeated experiments where they presented a shock at the tail (US) and a mild tactile stimulus at the tail (CS). These two stimuli were contiguous in time, that is, they occurred one after the other within a precise temporal interval. While in the first experiments Aplysia's gill-withdrawal reflex occurred only after the presentation of US, Kandel et al. found that, after repeated experiments, the CS caused the same reflex. The ability of the CS to cause the gill-withdrawal reflex seemed to depend on a process of presynaptic facilitation due to the neural pathway activated by the US: as a result of the activation of facilitator interneurons, the activity generated by US caused a greater facilitation of the sensory neurons responding to CS.

These results provided empirical evidence that the processes underlying sensitisation and classical conditioning were similar. Both depended on a process of presynaptic facilitation that, in the case of sensitisation, consisted in the greater production of neurotransmitter by siphon sensory neurons, while, in the case of classical conditioning, depended on the temporal pairings of US and CS, which, in turn, enhanced the presynapses of siphon sensory neurons.[19]

Why are these studies important to show the role that subpersonal explanations play in the conceptualisation of personal phenomena?

[19] For a more detailed discussion on the types of learning in Aplysia, see Gold & Stoljar (1999) and Hawkins & Kandel (1984).

Once we analyse carefully these experimental results, we recognise that the information gathered about the internal processes responsible for the various kinds of learning under consideration show that these types of learning are not as different in nature as previously supposed. In particular, the analysis of neurons, neural pathways and connections in the animal highlighted that classical conditioning is rather close in nature to one of the two kinds of learning usually considered simple learning: sensitisation. Here, certain subpersonal information offered evidence for a change in the constitutive nature of these personal-level categories, a result that wouldn't have been possible by observing the behaviour of the animal from a merely "outside perspective".

Cases like this one shouldn't be surprising to those who, differently from the autonomy theorists, believe that:

> "It is this sort of contribution that philosophers and neuroscientists expect neuroscience to offer in the future: a contribution to the way we think about the basic phenomena of the mind." (Gold & Stoljar, 1999, p. 826)

Tyler Burge in his *Origins of Objectivity* (2010) offers some additional compelling reasons for why we shouldn't think that constitutive information could only derive from personal-level explanations.

What is the nature of perception? Which are the constitutive conditions necessarily for a subject to perceive the world (i.e. to attribute physical features to specific physical particulars)? These are some of the questions Burge tackles in his book. However, differently from McDowell, Burge believes that scientific psychology can and does shed light on these questions:

> "Philosophy can help sharpen distinctions (such as that between perception and sensory discrimination, or between different conceptions of representation) that in scientific work are not as sharp as they might be. Science, in turn, provides applications, empirical content, and cases

> that enrich philosophical understanding and places limits on tenable philosophical positions." (*ibid.*, p. 26)

Burge distinguishes perception from mere sensory registration or sensory discrimination. In particular, he argues, perception is a form of objectified representation, that is, it is the capacity of a system to represent mind-independent features of the environment in a more or less accurate way:

> "Objective representation need not be derived, rationalised, validated by the individual. The most elementary forms of empirical objectivity are the products of conditions that the individual has no perspective on. Subindividual conditions are unconscious, automatic, relatively modular aspects of perceptual systems and belief forming system. Environmental conditions are twofold. They are the actual properties and relations in the environment that the individual interacts with and discriminates. And they are the patterns of causal relations between the environment and the individual's perceptual and cognitive capacities […]." (Burge, 2010, p. 24)

Sensory states are those states that register or encode information from the environment to contribute to an organism's fitness. A sensory state is, for instance, a state that encodes the information that in the environment there is a predator. This encoded information, then, initiates a process that affects the responses of the animal in such a way that they contribute to its fitness. When these states make their contribution, "it is not the accuracy *per se* that makes the contribution. The tendencies of the state to produce efficient responses to *needs* or, more precisely, tendencies to produce evolutionary fitness — not the veridical aspects of the state — make the contribution" (Burge, 2010, p. 302).

Organisms like bacteria and amoebae, for instance, discriminate particulars in the environment, such as light and heat. These discriminations, then, initiate a process that is good for the organism's fitness or survival.

A perceptual state is, instead, a representational state that can be veridical or non-veridical:

> "[…] a perceptual state is the type of state that it is partly by virtue of being a state that purports to pick out various particulars in a scene and to attribute to those purported particulars such attributes as being cube-shaped, being green, being in certain directions and at certain distances. If there are particulars causing the perceptual state in the right way and those particulars have the attributes that are attributed, the perceptual state is veridical." (*ibid.*, p. 308)

Perceptual accuracy doesn't necessarily or constitutively enhance biological success; rather, it is a constitutive part of representational success. Perception is therefore a type of veridical representation. To define veridical representations, subpersonal information from the sciences becomes necessary:

> "Where there is perception, there is sensory information registration. That is, where there is perception, there is functional, causally based, usually high, statistical correlation, between a type of state impacted by surface stimulation (and that encodes surface stimulation), on one hand, and a type of stimulation, on the other. Sensory information registration *per se* is not a type of perception […]. Perception is a sensory capacity for objectified representation." (*ibid.*, p. 317)

Burge offers the notion of objectified representation as a solution to the underdetermination problem: how can we explain the fact that an organism often represents the environment veridically given the fact that the encoded sensory information underdetermines environmental conditions?

> "To arrive at a representational state that privileges as *representatum* one of many possible environmental antecedents of the registration of sensory inputs, the system must have default settings, or a default range of possibilities for learning." (*ibid.*, p. 344)

According to Burge, *formation laws*, which describe the law-like regularities of the environment, help to overcome the underdetermination problem by transforming pure sensory states into representational states, whose contents are, at least, highly veridical. Burge describes the case of convergence as an example of transformation of sensory information in the visual system.

Convergence is one way the visual system determines the distance (i.e. the location) of an object in the environment. The lines of sight of the two eyes fixated on an object form an angle that is dependent on the distance between the subject and the object. A closer object corresponds to a wider angle, while a more distant object corresponds to a narrower angle. The fixation point and the middle point between the two eyes form a third angle and others can be identified using certain *geometrical principles* (see *ibid.*, pp. 348–349). Burge claims that:

> "Experiments have shown that visual systems rely on proximal information regarding [these] angles, together with the distance between the eyes to determine the distance and location of distal causes of proximal information." (*ibid.*)

Accordingly, a system can form representational states of an object in the environment because it can apply such principles. A perceptual representation then results from the conjunction of sensory information and geometrical principles and a constitutive explanation of perceptual experience needs to account for both. These two elements, when taken together, yield a form of objectified representation, which is what makes a state a perceptual state rather than something else. Note, however, that sensory information is clearly not a personal-level component and that formation laws are provided by scientific psychology.

If there are at least some cases where we need subpersonal information to explain what constitutes a personal-level phenomenon, then what about the autonomy of personal-level explanations over subpersonal-level explanations?

> "Perceptions and perceptual states that are attributed to an individual are always also attributable to the individual's perceptual system, a subsystem of an individual. Any visual perceptual state of an animal, for example, is also a state of the animal's visual system. Many processes that occur in perceptual systems, however, are not attributable to individuals. Transformations of sensory information into perceptions and transformations among perceptions are almost never attributable to the individual. The individual does not make them occur; they are not conscious or accessible to consciousness, they are not exercises of individual's central capabilities. But, necessarily and constitutively, *individuals* perceive. […] Individuals perceive as a result of perceptual states' being formed in their perceptual systems. Perceptual states are realizations of individuals' capacities." (Burge, 2010, p. 369)

The constitutive relation between perception and individual depends then on the fact that perceptual states constitutively figure in individual functions, that is, in fulfilling individual's needs and goals.

As I have shown, the distinction between personal and subpersonal explanations is not straightforward: subpersonal information can, and sometimes does, shed light on the nature of certain personal phenomena by saying something about what makes a certain personal-level state what it really is.

4.3.2 – The ideal of rationality

I will focus now on a second reason that is generally appealed to in order to argue for the autonomy of the personal-level of explanation with regard to the subpersonal-level of explanation: people's behaviour can be made intelligible once they are confronted with an ideal of rationality, and this ideal belongs only to the personal level.

Donald Davidson (1980) argues that personal explanations depend on what he calls the *ideal of rationality*, that is, explaining people's behaviour consists in making behaviour intelligible in light of the rational principles that govern them:

> "When we ask why someone acted as he did, we want to be provided with an interpretation. […] When we learn his reason, we have an interpretation, a new description of what he did which fits into a familiar picture. The picture certainly includes some of the agent's beliefs and attitudes." (Davidson, 1963, p. 691)

Along the same lines, Jennifer Hornsby claims that it is "a normative account that might show a person's doing something to be understandable" (Hornsby, 2000, p. 8).

Autonomists argue that a personal-level explanation of a behaviour is a normative explanation that cites the agent's reason (i.e. belief or attitude) for acting. Only a normative interpretation makes a person's behaviour intelligible in a way that "the redescription […] places the action in a wider social, economic, linguistic, or evaluative context" (Davidson, 1963, p. 691).

An action gets explained when it is placed within a wider pattern and we should accept the idea that personal-level phenomena cannot be further explained by going deeper than the level of persons. The fact that persons are rational agents is a sufficient reason to keep the personal-level and the subpersonal-level of explanation separate and independent from each other.

Let me try to uncover the nature and the role of these principles of rationality a bit more.

Principles of rationality regulate the explanation of agents' behaviour and, for this reason, they are normative rather than descriptive. This means that these principles describe how people *ought* to behave rather than how they actually behave.

A natural question might arise: do we explain a behaviour when we show that it is, or it is approximate to be, as rational as it ought to be? Consider the following quote:

> "[…] a person can have a reason for an action and perform the action, and yet this reason not be the reason why he did it. Central to the relation between a reason and an action it explains is the idea that an agent performed the action *because* he had the reason." (*ibid.*, p. 691)

Accordingly, an action is properly explained once the reason why a subject acted is identified; this reason really explains the action and it doesn't simply justify it. Nevertheless, when we redescribe an action we are not always in a position to do so because "what emerges, in the *ex post facto* atmosphere of explanation and justification, as *the* reason frequently was, to the agent at the time of action, one consideration among many, *a* reason" (*ibid.*, p. 697).

Normative explanations then run the risk of being nothing more than hermeneutic strategies in the sense that they explain intelligent behaviours by interpreting them as behaviours of rational agents (Bermudez, 2005, p. 42).

Rational interpretations, however, might not be good explanations. Imagine I see my friend Anna walking towards the department on a Saturday morning. I know that she is working on a chapter and that she has a close deadline. I then explain her behaviour by saying that she desires to finish the chapter and that she believes that it is easier for her to work in the office rather than at home. As a matter of fact, though, she simply passes by the department and walks straight ahead. My rational interpretation of Anna's behaviour is then not a good explanation of her behaviour although it makes it rational, appropriate and intelligible given what I know of what she wants (i.e. she desperately wants to finish her chapter) and of what she believes (i.e. she complained many times about how noisy her flat was during weekends). Indeed, an interpretation is a rational reconstruction of a behaviour that maximises its rationality: given that Anna needs to finish the chapter quickly and that her flat is always noisy during weekends, she ought to work in the office on Saturday morning. There are plenty of situations where we have a rational interpretation of someone's behaviour (or even of our own behaviour) that is not an explanation of it. Consider psychological and psychoanalytical literatures. These literatures are full of examples of people who believe they act for a certain reason, but that, actually, are moved by beliefs and desires that are consciously inaccessible to them. Some psychological studies (e.g. Nisbett & Wilson, 1977, 1978), for instance, show that we often construct stories that seem to make sense to us, and that sometimes we come up with reasons for our actions that, even if plausible, are demonstrably false. These cases are called confabulations (Hirstein, 2005): when we rationalise an action, we don't state the reasons that *actually* lead the agent to act in a certain way. A rational *ex post facto* description can be offered to make a subject's action intelligible even when we are ignorant about the course of events that led the subject to act in that way.

We are then left with a worry as to whether we should consider rational redescriptions to be genuine explanations rather than mere strategies to make behaviour intelligible. How can we justify personal and normative explanations? Is our only option for explaining behaviour that of asking people about their real reasons? If the answer to this question is affirmative, then a new problem arises since it has been long recognised that introspection is an unreliable method to establish the truth or falsity of explanatory claims (e.g. Nisbett & Wilson, 1977).

How can we then be sure that our rational redescription is an actual explanation of a certain behaviour?

> "[…] we can show […] that an action is rational if the acting subject can come up with a narrative that reconstructs how a given piece of behaviour fits into the agent's overall world view and character. What decides whether an action is based on reasons or not, then, is not whether these putative reasons have been always already in place, albeit subconsciously, but whether the behaviour in question can plausibly be made to fit into and cohere with the agent's overall system of desires,

> intentions, goals and values. […] It conflates the distinction between post hoc reasoning and confabulation, and makes episodes of explicit reason-giving which are accurate indistinguishable from cases in which subjects merely come up with a "coherent fiction." (Snow, 2006, p. 559)

If we remain at the personal level we run the risk of not being able to distinguish proper explanations from pseudo-explanations of cognitive behaviour. Personal-level explanations need to be supplemented by information from lower-level of analysis, both to uncover what is constitutive of personal phenomena, and to explain, rather than redescribe, cognitive behaviour.

Before turning to purely subpersonal-level explanations, let me summarise the arguments I endorsed to claim that purely personal-level explanations are not good explanations of mental phenomena:

- Subpersonal information can, and sometimes does, provide answers to constitutive why-questions about the nature of certain personal-level phenomena:
    - Kandel's studies on learning have shown that the nature of sensitisation and classical learning are more similar than expected
    - Burge's explanation of human perceptual experience needs to account for the capacity of objectified representation, which is understood as a conjunction of formation laws (defined by scientific psychology) and sensory data
- Purely normative explanations run the risk of being nothing more than hermeneutic strategies
    - They can make people's behaviour intelligible by placing it into a wider social, linguistic or evaluative context. Making people's behaviour intelligible, however, doesn't always mean explaining it, that is, identifying the real reasons for the behaviour

- Personal-level normative explanations can turn out to be pseudo-explanations of people's behaviour

4.4 – Subpersonal explanations

In contrast to personal-level explanations, which are typically couched in the vocabulary of folk psychology, subpersonal-level explanations come in various forms. Subpersonal explanations can be expressed in the vocabulary of cognitive psychology, in terms of brain components and their activities, or in terms of biological and chemical functions.

In the remainder of this chapter, I analyse two subpersonal views with regard to mental phenomena: eliminativism and reductionism. I first consider Paul Churchland's eliminativism and then John Bickle's reductionism. I address the arguments for the goodness of explanations of mental phenomena couched purely in subpersonal neuroscientific terms, and I discuss whether these arguments are justified or not.

4.4.1 – Churchland's eliminativism

> "Our vocabulary of propositional attitudes should be viewed as a simplification of the underlying multidimensional-reality, a conceptual framework whose predictive and explanatory utility indicates not its accuracy, but the extent to which it abstracts away from and compresses the underlying complexity." (Churchland, 1981, p. 129)

Churchland's famous paper "Eliminative Materialism and the Propositional Attitudes" is a manifesto for those who take an eliminativist stance towards folk psychology and believe that it should be replaced by neuroscience.

The paper sets up a naturalistic methodological position: a person interested in understanding mental phenomena needs to recognise that the mind is a physical object and that a proper explanation of mental phenomena should place them within the natural realm.

Given this methodology, Churchland's first and main critique of the goodness of folk-psychological explanations is that folk psychology doesn't pick out anything real. Since mental states are not physical states, they are empty notions and, because folk psychology is based on mental states, folk psychology is false. This, in a nutshell, is Churchland's eliminativist argument.

Churchland's metaphysical position is materialism: mental states, like beliefs and desires, are non-physical entities. Only physical entities really exist in the natural world, hence we should eliminate the folk notions from the vocabulary we adopt to explain mental phenomena. There is nothing more to the understanding of the mind than understanding the brain.

The starting point in this debate consists in treating folk psychology as a theory. If folk psychology is a theory, then "it is at least an abstract possibility that its principles are radically false and its ontology is an illusion" (*ibid.*, p. 72).

According to Churchland, however, for folk psychology this is not an abstract, but a very concrete possibility. Folk psychology is unable to explain many different cognitive phenomena, such as mental illnesses, creative imagination, individual intelligence differences, sleep, the ability to perform motor actions, memory, perceptual illusions and also learning. Failures in accounting for these phenomena show that folk psychology cannot explain all of mental life.

According to Churchland, the status of folk psychology should be reconsidered not only in light of its explanatory failures, but also in light of its stagnancy: folk psychology "is no part of [the] growing synthesis" (*ibid.*, p. 75) because it hasn't progressed since the ancient Greek times and, in contrast to other sciences such as evolutionary theory, biology, neuroscience and physiology, folk psychology's intentional concepts haven't evolved since. Considering these and other features, folk psychology is an appealing candidate for elimination.

There are two immediate objections to this position, which Churchland considers in turn.

A first objection draws on the normativity of personal-level folk-psychological explanations: folk psychology is a normative theory and not an empirical theory. According to autonomy theorists, beliefs, desires and other mental states do not cause cognitive behaviour; rather, they are the reasons for them. Only empirical causal theories can be falsified by empirical data. Folk psychology is not an empirical theory, hence it cannot be empirically falsified.

Churchland's counter-argument consists in showing that the normative dimension of folk psychology should be understood from a naturalistic perspective. He argues that the existence of certain regularities among propositional attitudes doesn't show that there is something normative about these regularities. In the same way in which we don't attribute a normative character to the gas law's regularities, we shouldn't give a normative gloss to the folk-psychological ones. Indeed:

> "[…] logical relations between propositions are as much an objective matter of abstract fact as are arithmetical relations between numbers. […] A normative dimension enters only because we happen to *value* most of the patterns ascribed by FP." (*ibid.*, p. 82)

The second line of critique that Churchland considers has to do with the abstract character of folk psychology.

Functionalists, like Fodor, claim that folk psychology should be understood as an abstract theory that doesn't need to be informed by the nature of the brain. According to this view, folk psychology refers to states that are internal and that are characterised by what they do, by their reciprocal relations and by how they are connected to the inputs and the outputs of a given cognitive process. Given that each psychological state can be realised in multiple ways in the brain, we should maintain that a functional characterisation plays an important role in explaining people’s behaviour even if it doesn’t mirror the structure and functioning of individual brains.

Churchland’s reply to this objection stresses the dependence of functional characterisations on implementational details. He argues that, if we highlight the functional role of mental states as a way of claiming that explanations of mental phenomena are independent from knowledge of the brain, then we are on the wrong track because this would amount to say that functional categories could never be false, which would be absurd. On the contrary, Churchland strongly believes that a good theory has to leave room for the possibility that the categories it identifies might be wrong. The explanatory failures of folk psychology make it a real possibility or at least provide reasons for remaining skeptical about its truth.

Churchland compares the folk-psychological functional concepts with the concepts employed in the past by alchemy. According to Churchland, in the same way in which alchemy was replaced by modern chemistry, folk psychology will be replaced by modern neuroscience:

> “[…] the correct account of cognition, whether functionalist or naturalistic, will bear about as much resemblance to [folk psychology] as modern chemistry bears to four-spirit alchemy.” (*ibid.*, p. 82)

To sum up, Churchland argues that:

- Mental states are empty notions that do not belong to the natural world
- Folk psychology can’t explain many cognitive phenomena, so it is not so predictive as we normally think it is
- Folk psychology is stagnant and unproductive, and these are not features of good theories

In the next section I will claim that Churchland’s arguments don’t justify the endorsement of eliminativism.

4.4.1.1 – Why not eliminativism?

Churchland claims that folk-psychological concepts fail to refer. Given that mental states are the main components of the theory of folk psychology and that they are not real, folk psychology is false. Neuroscience, instead, identifies real entities and processes in the brain, hence it will provide proper explanations of mental phenomena.

As others have already noted (e.g. Gold & Stoljar, 1999), without a clear empirical confirmation for this claim, eliminativism risks remaining only an interesting but hypothetical position. Indeed, there are cases where folk psychology *does* not only play an important heuristic role in driving further research into the nature of cognitive phenomena (i.e. having a clear behavioural description of a cognitive behaviour is important to look for the possible neural mechanisms underlying it), but also a constitutive role in their explanations.

Consider the example of learning that I have discussed above with respect to the distinction between constitutive and enabling conditions. The first thing to notice is that Kandel's low-level theory did not develop in a vacuum. Rather, the theory makes use of some psychological works, and, in particular, it draws on the psychological model of classical conditioning (e.g. Rescorla, 1968). In addition to this, the example shows, on the one hand, that certain findings at the lower-levels can shed light on important aspects of the explanandum phenomena that might have gone unnoticed at the personal-level of analysis, and, on the other hand, that certain personal-level notions figure in explanations of cognitive phenomena. Let me elaborate on this last point.

From one perspective, we could say that Kandel and Schwartz (e.g. 1982) offer a purely neurobiologial explanation of classical conditioning in Aplysia according to which conditioning is a process where the contiguity between the conditioned stimulus (CS) and the unconditioned stimulus (US) transfers the response from the former to the latter. This characterisation, however, doesn't seem to properly explain the data.

In Rescorla's first experimental setting (1968), a rat was exposed to a tone during a 2 minute interval (CS) and to an electric shock (US) at a later time within the same interval. In the second experimental setting, the electric shock occurred when the tone was presented. The results showed that only the rat that participated in the second experiment learned the association between the tone and the shock.

Interestingly, these results couldn't be explained by making reference to the notion of contiguity as the accepted view of classical conditioning suggested (i.e. two stimuli occurred within a certain temporal interval — 2 minutes) because the CS and the US satisfied the contiguity requirement in both experiments. Given that conditioning happened only in the second case, something more was needed to make sense of these results.

Rescorla claimed that the extra-ingredient in the second experiment was *information*: conditioning learning could be explained by making reference to CS and US, but also to the notion of *information*, that is, by saying that the tone was providing information about the coming of the shock:

> "[…] in order to explain this effect […] one needs to appeal to a notion of information that is richer than the notion of low-level mechanical process in which the control over a response is passed from one stimulus to another." (Gold & Stoljar, 1999, p. 824)

Indeed, learning is a process where the difference between the actual state of the world and the organism's representation of that state is reduced, which is a way to say that the organism learns when it can reduce its *surprise*.

It is hard to see how the concept of synaptic plasticity can capture the conceptual complexity of the notions of information and surprise. These are personal-level notions that characterise the behaviour of the whole organism: it is the organism that can use information and that reduces surprise via certain synaptic mechanisms. Certain personal-level notions, then, remain present even in lower-level theories:

> "No matter how much we know about the causal intricacies at the molecular level, there is always this further question to be asked, namely, what are the causal mechanisms for, what is it that these mechanisms […] are supposed to accomplish?" (de Jong & Schouten, 2005, p. 480)

Can Churchland's eliminativist claims convince us that folk-psychological personal-level notions and explanations should be eliminated? My answer has been negative and here are the reasons I provided:

1. Folk psychology is useful in driving further research and in setting up experiments to understand the nature of mental phenomena
2. Certain personal-level concepts can be part of successful explanations of cognitive phenomena (e.g. the notion of information and not the lower-level notion of contiguity is required to explain conditioning learning)

In the next section I will consider John Bickle's subpersonal account. He argues that folk psychology cannot explain mental phenomena because it plays a purely heuristic role. His aim is to show that the causal role of high-level folk-psychological explanations of a range of mental phenomena is dropped once lower-level explanations of the same phenomena are found.

4.4.2 – Bickle's reductionism

John Bickle (e.g. 2003, 2006, 2007) argues in favour of reductionism within the philosophy of mind. His main claim is that psychological concepts and kinds should be reduced to molecular-biological mechanisms and pathways, and he is confident that the reduction of mind to molecules is forthcoming in contemporary neuroscience. He admits that philosophers have been unable to notice these reductions and motivates this failure by saying that they were too focused on the relationship between higher-level and lower-level explanations and too little on advances in molecular and cellular neuroscience. In particular, Bickle says, philosophers haven't analysed the neuroscientific experimental practices that "bridge molecular pathway levels *directly*; and this practices are common to all recent empirical successes" (Bickle, 2006, p. 414).

He argues that lower-level explanations can explain mental phenomena in terms of their observable manifestations and their dependencies on molecular modifications and interventions. Certain molecular interventions, he claims, clearly show that behaviours and behavioural changes can be properly accounted in terms of changes in molecular activity in the brain. In particular, we are justified in concluding that certain molecular mechanisms are the mechanisms of, for instance, memory consolidation when we can:

> "[…] *intervene causally* at increasingly "lower" levels of biological organization in animal models and then to track the specific effects of these interventions on behaviour in widely accepted protocols for the cognitive phenomenon under investigation." (Bickle, 2007, p. 230)

Within Bickle's account, "one only claims a successful *explanation*, a successful *search for a cellular or molecular mechanism*, or a successful *reduction*, of a psychological kind when one successfully intervenes at the lower level and then measures statistically-significant behavioural difference" (Bickle, 2006, p. 420).

If the interventionist strategy works by producing the expected behavioural change, then, according to Bickle, we can confidently say that we have explained the behavioural data *directly*, with no need to make use of intermediate levels (e.g. cognitive neuroscience, information processing, and so on).

Direct explanations are distinguishable features of Bickle's reductionism with regard to the more classical interlevel reductionism: interlevels are simply not necessarily to reduce the mental phenomena to the molecular level. The role of higher-level theorizing is solely heuristic in the sense that "once [they] have served their heuristic function — once the appropriate higher-level tool, theoretical assumptions, and experimental results have identified candidate cellular or molecular mechanisms scientifically — they give way to the strategy of 'intervening cellularly/molecularly and tracking behaviourally'" (*ibid.*, p. 428).

Bickle's reductionism differs also from other possible readings of "reduction" in that it doesn't require, or assume, that it is possible to explain cognitive phenomena in terms of lower-level laws or generalisations. He argues that in cellular and molecular neuroscience there are very few explanations that appeal to laws or generalisations: molecular biologists know how specific molecules interact in specific contexts, but they don't provide explanatory generalisations for these processes.

In brief, Bickle's reductionist arguments can be summarised in the following way:

- Philosophers have underestimated the importance of experiments in molecular neuroscience because they were too focused on understanding the nature of interlevel relations
- Molecular and cellular neuroscience show that behaviours and behavioural changes can be explained in terms of molecular activities: once we have operationalised a behaviour, we can intervene on certain molecular variables we believe are responsible for the behaviour and then measure the effects
- Molecular neuroscience shows that behaviours can be explained *directly* with no need to rely on intermediate levels of analysis

A close look at neuroscientific findings, however, seems to suggest that the direct link between psychology and molecular neuroscience that Bickle addresses is not so obvious.

A direct molecular explanation of a mental phenomenon depends, in Bickle's account, on how researchers operationalise it. However, practice in science shows that different researchers tend to operationalise the same behaviour in different ways, thus yielding different explanations of the underlying molecular mechanisms responsible for it. These differences seem to undermine Bickle's claim that contemporary neuroscience has already provided examples of reduction.

In what follows I will analyse the consequences of Bickle's reductionism by drawing heavily on Jacqueline Sullivan's descriptions of the multiplicity of experimental protocols in neuroscience (2009).

4.4.2.1 – Do we really have reductionist explanations?

To understand the consequences of Bickle's position, Sullivan discusses one of Bickle's case studies: social recognition memory in mice and, in particular, the role that a specific protein (CREB) plays in it. Social recognition memory is generally operationalised in terms of the ability to recognise another individual after an initial interaction with it.

Bickle considers the experiments run by Kogan et al. (2000) as examples of reduction. Kogan and colleagues intervened on CREB to have mutant mice that were deficient in two isoforms of CREB and that had a reduced amount of CREB in their brains. They trained a group of mutants with a group of normal mice using a modified behavioural protocol associated with a previously developed learning paradigm. In their experiments, a normal mouse was placed in a cage with a mutant adult or with a normal adult already habituated to the cage for 15 minutes for a first interaction of 2 minutes. The experiments were then followed by 24 hours delay. After this period, the adult mouse was observed while socially investigating the new mouse. The types of behaviour that were considered cases of social investigations included: direct contact with the new mouse while inspecting any part of the body, sniffing of the mouth, of the ears, of the tail, of the ano-genital area and close following of the new mouse (within 1 cm).

Kogan and colleagues found that, differently from normal mice, the mutants examined the new mouse in the same way even after the 24 hours intervals. They interpreted these results as signs of the inability of the mutant mice to encode in long-term memory the information necessary for the recognition task. These failures were associated to the CREB mutation.

Examples like this one show that it is possible to track behavioural changes by intervening on molecular variables; Bickle calls this methodology *reduction-in-practice*.

Despite the appeal of Bickle's proposal, there are at least two problems with his interpretation of Kogan's studies. A first problem is related to the claim that these studies are examples of reduction-in-practice of social recognition memory; a second problem has to do with the more general reductive claim he endorses.

Concerning the first problem and in contrast to what Bickle claims, neuroscientific experiments of the same phenomenon (e.g. social recognition memory) are quite diverse. This diversity makes experimental results difficult to compare and integrate:

> "Reduction-in-practice is something that occurs in an individual experimental laboratory when an investigator operationalizes a psychological function […] by developing a protocol that specifies how to produce that function […] and detect when it occurs, by reference to observable changes in behaviour […]." (Sullivan, 2009, p. 517)

The experimental protocols can vary in relation to the duration of the exposure of the mutant mouse to the new mouse, in the interval duration or in the behavioural features that are supposed to mark the acquired capacity of social recognition. In most of the cases, different molecular mechanisms can be used to directly explain the behavioural data of that particular experiment, in that particular laboratory, obtained by following a specific experimental protocol. Relatedly, it is still an open question whether the operationalisations are appropriate given that researchers simply "*assume* that the operationalizations […] are actually indicative of the function of interest" (*ibid.*, p. 518).

The second concern is about Bickle's more general reductionist claims. Bickle takes the ability of molecular neuroscience to directly explain *certain* cognitive phenomena as a sign of its ability to explain *all* cognitive phenomena. This follows from his general claim that higher-level explanations and notions have only a heuristic role in the discovery of the proper molecular explanations. Nevertheless, given the difficulty that the reductionist methodology already shows in relation to a single phenomenon and the fact that "at best what Bickle has achieved by appeal to experiments in molecular and cellular cognition is a case for many local "within-lab" reductions" (*ibid.*, p. 519), the conclusion that the future of neuroscience will provide reductions for all cognitive phenomena in non-humans and in humans is unwarranted.

While Bickle believes that direct explanations can be offered by looking at practice in science, and in particular in molecular neuroscience, I have discussed some drawbacks of his account. Once we recognise that, on the one hand, it is not easy to describe the relation between a certain mental phenomenon and its underlying molecular mechanism, and that, on the other hand, lower-levels explanations do often bear the stamp of the higher-level ones (see for instance Kandel's study above), we are forced to admit that higher-level notions and theories are not superfluous or purely heuristic as Bickle (and Churchland) wants to make us believe.

To sum up, a reductionist account doesn't succeed in undermining the importance of personal-level concepts and explanations because:

- Reduction of mind to molecules is not forthcoming in neuroscience and the reduction-in-practice example of social recognition memory that Bickle addresses doesn't constitute a case of real reduction:

  - The way in which scientists operationalise an ability varies and, as a consequence, different molecular mechanisms can be discovered for the same capacity
- Even once we have a reduction of a mental phenomenon at the molecular level, it doesn't follow that all mental phenomena will reduce to molecular phenomena *directly*

4.5 – Conclusion

Three different answers to the interface problem have been examined in this chapter.

Autonomist theorists provide a first answer by claiming that personal-level explanations are autonomous from subpersonal-level explanations. They argue that the two types of explanation aim at accounting for different phenomena: personal-level explanations answer constitutive why-questions, while subpersonal explanations answer enabling how-questions. I claimed that their arguments don't succeed for two main reasons: (i) subpersonal information can, and sometimes does, provide answers to constitutive questions; (ii) a normative explanation of a phenomenon runs the risk of being purely hermeneutic, but not explanatory.

I then analysed the materialistic approach according to which folk psychology is a false theory and its concepts are empty concepts that should be replaced by neuroscientific ones. I argued against this proposal by claiming that certain personal-level concepts can figure in successful explanations of mental phenomena and that folk psychology is not false but needs to be enhanced.

The third and last position that I have considered is a reductionist position according to which mind should be reduced to molecules and neuroscience is already providing examples of reduction-in-practice. I claimed that neuroscientific practice is very diverse and that this limits even the possibility of local reductions, that is, reductions of single mental phenomena. As a consequence, the claim that all mental phenomena will get reduced to molecular mechanisms is unwarranted.

The analysis of the distinction between personal- and subpersonal-levels of explanation that I have just offered suggests that both levels are needed to adequately explain cognitive behaviour.

Chapter 5 - Jose Luis Bermudez on Rationality and Reasoning

5.1 – Introduction

In the previous chapters, I argued that adequate explanations of why systems perform cognitive behaviour require the employment of the notion of representation. I also claimed that normative personal-level explanations can make behaviours intelligible but cannot provide good explanations of them and I suggested that an identification of the subpersonal mechanisms underlying cognitive capacities is necessary to justify the validity of personal-level claims.

In this chapter, I will consider a different take on personal-level folk-psychological explanations by engaging with Jose Luis Bermudez's account of rationality (2000, 2003, 2009).

I will first highlight the differences between his notion of rationality and that of autonomy theorists and I will then examine his arguments in favour of three different levels of rationality. I will argue that the two criteria for rational behaviour that Bermudez identifies (i.e. the behaviour results from a range of alternatives and the behaviour matches some normative standards) are inadequate to understand the nature of rational behaviour. I will show how adequate explanations of rational behaviour are possible only once external behavioural criteria are complemented with internal mechanistic ones. Details about how information is encoded and manipulated inside the brain, I will claim, are essential to confirm or disconfirm our hypotheses about the role and nature of reasoning processes and, ultimately, to evaluate our hypotheses about how rationality is naturally possible. Throughout the chapter, I will also distinguish between "objective" and "subjective" utility, between adaptive and individual goals, and between instrumental components (and selection processes) and instrumental beliefs (and decision processes).

5.2 – Personal-level explanations

Before engaging with Bermudez's account of rationality and rational behaviours, let me briefly recall the main features of personal-level normative explanations.[20]

According to folk psychology in general and to the so-called autonomy theory in particular, an agent's behaviour is explained once its responsible mental states are identified. These states, which are characterised by their propositional content, are related via rationalising connections. An agent whose behaviour can be explained in this way is a rational agent.

While propositional attitudes are seen as the reasons motivating agents' action, rational constraints allow to cut down the number of possible variations associated to an agent's *psychological profile* (i.e. the combination of the agent's beliefs and desires — Bermudez, 2009) and her consequent action. In particular, a behaviour is considered rational when it maximises a certain utility with respect to an agent's goal and current environment. A personal-level folk-psychological explanation of a behaviour *a* typically takes the following form: if an agent *s* desires *p* and knows that by doing *a* she will get *p*, then *s* will, *ceteris paribus*, do *a*.

[20] For more details on personal-level normative explanations, see chapter 4.

For a folk psychologist, rationality is linked to the ability to entertain a conscious inferential process by constructing an argument from premises (beliefs and desires) to conclusions (behaviour). Recently, however, there have been some interesting suggestions to reconsider the argument-form attributed to these processes. Hugo Mercier and Dan Sperber (2011) claim that the term "inference", as it is normally used in psychology, refers to processes where new mental states are generated from previously held mental states. When inference is understood in these terms, the production of new beliefs on the basis of previous beliefs, the production of expectations on the basis of perception and the elaboration of plans on the basis of beliefs and desires become all cases of inference. If inference doesn't necessarily need to be associated with a deliberate and conscious process and with the consequent ability to construct a valid argument, then it can be seen as a building block of not only conceptual thinking, but also of perceptual and motor processes.[21]

As I will show throughout the chapter, Bermudez (2003) addresses a similar claim. Although he stresses the importance of normative rational explanations, he believes that a behaviour can be internally and externally rational even if it does not result from a conscious inferential process. According to him, rationality can be attributed to non-linguistic creatures as well as to low-level behaviours (e.g. perception).

5.3 – Bermudez on rationality

Rationality is usually considered a normative notion that defines the way in which an agent ought to behave, while reasoning is a descriptive notion that specifies the

[21] The possibility of considering perceptual and motor processes already as instances of rationality has important consequences that will be considered later on in the thesis (see chapters 6 and 7).

process performed by an agent to behave rationally. Rationality and reasoning are both notions that belong to the folk-psychological vocabulary.

A folk-psychological account of rationality considers a behaviour rational (and the agent performing that behaviour a rational agent) if it results from a conscious inferential process.

Bermudez draws on folk psychology, but claims that a behaviour is rational as long as: (i) it results from a range of alternatives; (ii) it matches some normative standards. Organisms or agents can engage in inferential tasks even by relying on simple rules and heuristics.

Bermudez suggests three different levels of rationality. One level is only externally rational, while the other two are also internally rational. In his own words:

> "[The] assessments of internal rationality are relative to an agent's doxastic and motivational states, taking those states as given, while assessments of external rationality include assessments of the doxastic states underlying the action. To say that an action is externally rational is to say that it is in some sense appropriate to the circumstances in which it is performed, where those circumstances include the agent's motivational states—with different theories of external rationality interpreting the type of appropriateness involved here in different ways. The internal rationalizing connection between beliefs, desires, and actions allows the attribution of thoughts and desires to be genuinely explanatory. Beliefs and desires cause behaviour qua beliefs and desires (that is to say, in virtue of their content) because their contents rationally dictate a single course of action—or a limited number of possible courses of action. In the absence of such a rationalizing connection, there would be no reason why a belief-desire pair with those particular contents should cause that particular action." (Bermudez, 2003, p. 110)

I quoted this passage at full length because it clarifies the distinction between external and internal rationality. It also shows that Bermudez somehow equates rationality with appropriateness in behaviour: we can assess if a behaviour of an agent is externally rational, he says, by judging whether it is appropriate given the current circumstances and the goal of the agent. To judge whether a behaviour is also internally rational, instead, we need to focus on the nature of the intermediary processes that lead from desire/goal to behaviour. The components of these intermediary processes have to be mental states or propositional attitudes, and they have to be linked by rationalising connections. These connections are equivalent to the rationality constraints I mentioned above: rationalising connections among beliefs, desires and goals cut down the number of possible rational courses of action. In Bermudez's account, the rationalising connections don't simply cut down the number of possible rational actions, but they also usually dictate, depending on the content of mental states, a single rational action. More generally, an organism can behave rationally if it has a space of possible alternative actions available.

Following the distinction between internal and external rationality, Bermudez identifies three rational levels: level 0, level 1 and level 2 rationality.[22]

5.3.1 – Level 0 rationality

Level 0 rationality characterises those behaviours that, according to Bermudez, do not require a psychological explanation. These are all sorts of automatic stimulus-response behaviours, from reflexes to innate mechanisms, that don't traffic in representations mediating inputs and outputs.

Consider the following level 0 rational behaviour. Imagine you are in a given context *x* at time t_1. There is a flame next to you and you decide to move your fingers

[22] In *Thinking without words*, Bermudez considers the possibility that also non-linguistic creatures can perform rational behaviours. I believe that the same reasons he uses to address and justify this possibility can also be applied to humans, especially when the phenomena to be explained are low-level ones.

close to it. As soon as you reach the fire, you immediately move your fingers away. According to Bermudez, your behaviour is an instance of level 0 rational behaviour: it is automatic, it might have evolved to preserve your species and it doesn't involve any decision process.

All members of a same species perform the same level 0 rational behaviour given appropriate environmental conditions. For this reason, Bermudez calls these behaviours "dispositions": a level 0 rational behaviour is a disposition of a species as a whole, rather than of any individual organism.

These dispositions are sort of genetic dispositions and they are called rational because, by performing in conformity to them, single organisms can successfully perform in their environment. If the organism had acted in conformity to a different disposition, its behaviour would have been less rational.

The behaviour described above doesn't result from a deliberate choice: if we get close to the fire, we move away very quickly, without considering any other possible course of action. Nevertheless, Bermudez affirms that even this simple kind of behaviour is an instance of rationality: it is rational because it is adaptive and because it is appropriate with respect to the environment and to the goal of the organism.

Another interesting example is that of foraging behaviour. According to Bermudez, foraging behaviour is a level 0 rational behaviour because it is externally rational, that is, it allows animals to maximise their search for food. Indeed, a specific foraging behaviour can be explained in terms of an optimal foraging theory: it is one of a set of possible foraging behaviours and it can be compared with a normative standard (i.e. the maximisation of energy that the animal can obtain from food).[23]

To summarise, level 0 rationality can be assessed in relation to a behaviour on the basis of external criteria of rationality, which means that a behaviour is rational because its outcome matches some normative standard. Considerations about the

[23] For more details on level 0 rationality, see Bermudez (2003, pp. 116–120).

underlying processes are not relevant. In particular, a level 0 rational behaviour is rational even if it doesn't result from any internal operation over representations.

5.3.2 – Level 1 rationality

Level 1 rationality applies to those cases where there is a set of possible behaviours available *here and now* to a specific organism, among which just one is selected, and where the selection does not result from a decision process.

Consider the case of an animal that is confronting a dangerous animal and that has to choose one of two possible behaviours: fight or flee. Bermudez argues that, although we wouldn't say that the animal is engaging in any genuine decision-making process, "there is a clear sense in which one of the two possible courses of action could be more rational than the other" (Bermudez, 2003, p. 121).

One of the two behaviours carries a greater advantage for the animal because it maximises a certain expected utility with respect to its goal. According to Bermudez's account, if the animal's behaviour maximises the expected utility, then the behaviour is rational. It is rational although it doesn't result from a decision-making process.

To clarify the difference between selection and decision processes, Bermudez draws on the Gibsonian theory of affordances (Gibson, 1979): in a selection process, an organism is able to select the most appropriate behaviour by relying on direct perception. In the example above, the animal selects fight instead of flee because it just "sees" that fighting is the most appropriate course of action. This means that: (i) the animal performs rationally by relying on vision alone; (ii) in cases of level 1 rational behaviour, perception is direct (i.e. the animal not only perceives that something is in its environment – a dangerous animal – but also what it can do in response to it); (iii) the content of the animal's perception corresponds to both the presence of a dangerous animal and the possible actions.

Bermudez argues that an organism doesn't need to represent affordances in any complex way because they stand for possibilities of actions that are already part of the content of perception itself. The comparison of these simple representations doesn't require any decision-making process, that is, it doesn't need to follow any step-wise procedure. Consider Fodor's model of practical decision-making (1975); this model consists in different steps:

1- The organism is in a certain situation S
2- The organism believes that there is a set of possible courses of actions A_1, A_2, A_3, …, A_n available in that situation S
3- The organism predicts which consequence C would probably follow from selecting and performing each of the possible courses of actions (i.e. action A_1 in S will probably yield consequence C_1, and so on)
4- The consequences are ordered in accordance to their preferences
5- The organism will choose the action with respect to the probability and preference of its consequence in situation S

An animal that is confronting a dangerous animal does not apply this form of decision-making process; rather, it relies only on the content of its direct perceived environment and affordances. The animal only "sees" what to do: it does not need to consider all the possible courses of actions, calculate the probability of their outcomes in a given context, and then perform the action that, according to this calculation, is likely to maximise a certain kind of expected utility.

Given this characterisation of level 1 behaviour, there is a clear sense in which affordances differ from mental states (e.g. beliefs and desires) within Bermudez's account. To highlight such differences, Bermudez calls affordances *instrumental components*. For an animal to compare the action of fighting with the action of fleeing, it only needs "representations of actions" (*ibid.*, p. 123), which are not very complex and, he says, can be understood at a purely perceptual level. These representations, or affordances, enable the animal to behave rationally by selecting the action that maximises the current utility.

The range of behaviours is here assessed with respect to a specific organism in its ontogeny (i.e. organism's development) and not phylogeny (i.e. species' evolution), as it was for level 0 rationality, and it is closely dependent on the here and now of the organism's interaction with its perceived environment.

Capacities that are typically considered to be non-cognitive and non-rational, such as perceptual abilities, are examples of level 1 behaviours. Indeed, Bermudez offers an account of rationality that can embrace high-level but also low-level processes. Perception is rational, he claims, because it enables the organism to select the action that maximises a certain kind of utility. Perception, in particular direct perception, is the process through which the animal chooses the "best" and most context-appropriate behaviour.

Although the range of alternatives is limited to those afforded by the environment, this limitation, Bermudez says, does not affect the applicability of the notion of rationality; rationality can be applied whenever the behaviour is selected within a range of other possible behaviours, no matter how extensive and numerous that range is.

5.3.3 – Level 2 rationality

Level 2 rationality is different from the previous two types of rationality in that it identifies behaviours that result from genuine decision processes. For a process to be a decision process it needs to select a particular course of action on "consequence-sensitive grounds" (Bermudez, 2003, p. 124).

For a behaviour to be selected on consequence-sensitive grounds, it means that it is selected for a reason. The reason involved is associated to the probability of the action's outcome: an organism decides when it has reasons to assert, given its goal and current context, that a specific course of action can, better than others, maximise its expected utility.

Bermudez calls these reasons *instrumental beliefs* since they are representations of the contingencies that hold between a given action and its expected outcome and because they inform the cognitive system about how to behave to satisfy its goal in a given context. In this sense "to say that an action is being performed on consequence-sensitive grounds implies far more than it's simply being performed because of its consequence" (*ibid.*). For a behaviour to be assessed at level 2 rationality, then, the organism needs to have a goal and instrumental beliefs (i.e. reasons to act in a certain way in order to satisfy a goal).

Level 2 rationality shares some features of personal-level explanations of rational behaviour. It considers an organism rational when it behaves in accordance with reasons, which link its beliefs and desires in rational and appropriate ways with respect to a given environment. An action is a level 2 rational action when there is a match between the action and the organism's background beliefs.

In contrast to the folk-psychological account of rationality, however, Bermudez believes that representations of contingencies and their comparison do not necessarily have to be thought in terms of classical inferences.

Consider tool manufacture and tool-using behaviours. According to Bermudez, they are clear examples of level 2 rational behaviours because they depend on representations of contingencies between actions and their outcomes. Wild chimpanzees make wands for dipping into ants' nests in one way and wands for dipping into termite nests in another way. The tool construction techniques involved in these two cases are different both in terms of the materials used and in terms of the processes adopted. Moreover, chimpanzees often decide how to construct their tools far away and well in advance with respect to where they are actually going to use them. This suggests that chimpanzees can predict the future and act in accordance with their predictions: tools that will be used to catch ants are different from those that will be used to catch termites.

Although some people (Gould & Gould, 1998) think that even these behaviours are innate, as level 0 rational behaviours, Bermudez believes that they are good examples of behaviours that result from thinking processes, in which representations of contingencies are used.

Accordingly, instrumental beliefs and genuine decision processes enter the picture only when an animal can form a belief about the consequences that an action is likely to have. In particular, one of these two conditions needs to be met:

- The organism should not be able to perceive directly the goal of the action
- The organism should not immediately perceive that a certain action would yield the desired result (an animal can directly perceive a goal and still not know how to obtain it until it forms an appropriate instrumental belief for it)

If neither of the above conditions is met, the organism does not need to rely on explicit beliefs but only on instrumental components, as it is in the cases of level 1 or level 0 rationality.

Bermudez maintains that the difference between level 1 and level 2 rational behaviours lies in the fact that in the first case instrumental components are the contents of perception already, while in the second case the organism has to create separate instrumental beliefs. This distinction becomes more evident when operational criteria are adopted.

5.3.4 – The explanatory role of operational criteria

Bermudez suggests that we can rely on operational criteria to prove that a behaviour is an instance of level 2 and not of level 1 rationality.

In operational terms, Bermudez says, a level 2 rational behaviour is typical of an organism that can modify its behaviour once contingencies in the environment change. If a given contingency between an action and its expected outcome ceases to exist, a level 2 rational animal should stop performing on the basis of that contingency and select a more context-appropriate behaviour. If, on the other hand, an animal persists in behaving in a certain way even in the absence of a given contingency, then its behaviour belongs to level 1 and not to level 2 domain.

An example is the Rescorla and Skucy's studies on rats (1969). This series of studies showed that rats trained to press a lever to get food ceased to press the lever if the food started to be delivered in correspondence to different circumstances. Rats were able to recognise that there was a contingency between lever-pressing and food delivery, and also that, from a certain time, that contingency didn't hold anymore. According to Bermudez, rats performed level 2 rational behaviours.

Other examples of level 2 rational behaviours are those concerning actions that clearly go beyond the current available affordances. If an animal is able to decide not to act on the affordances that are directly available, it means that the animal is deciding on the basis of some kind of instrumental belief. This belief is such that the animal decides to take a different course of action than the ones afforded because the selected action is considered to yield the maximum expected utility with respect to its goal. In these cases, the instrumental belief (i.e. the representation of a contingency) informs the animal that if it performs a certain action it will probably get the maximum expected reward available.

Bermudez claims that these minimal operational criteria show when a behaviour is rational in terms of level 2 rationality and when it is rational in the sense of level 1 rationality.

5.4 – Critical discussion

5.4.1 – Level 0 *non*-rational behaviours

I start discussing level 0 rationality that, according to Bermudez, applies to cases in which:

- We are considering an automatic stimulus-response behaviour: given the same input, the organism always responds with the same behavioural output
- Level 0 rational behaviour doesn't result from the manipulation of intermediary internal representations
- Level 0 rational behaviour is a rational "genetic disposition": the range of possible alternatives does not characterise a single organism, but a species as a whole that has adapted to follow rational rather than less rational dispositions with respect to the environment
- Rationality at this level is external rationality as the focus is on behavioural outcomes
- Rationality can be applied to a behaviour when there is a normative standard (i.e. the maximisation of some expected utility) against which it can be compared

In what follows I will argue that level 0 rational behaviours are not instances of rationality by providing the following reasons:

i. Level 0 rational behaviours do not require a psychological explanation
ii. Level 0 rational behaviours maximise an objective kind of utility with respect to adaptive goals (i.e. survival and reproduction). This maximisation can be accounted for in terms of natural selection and adaptation

For point (i), I argue that level 0 rational behaviours don't require psychological explanations. As Bermudez himself affirms, psychological explanations are required only if behaviours cannot be predicted purely on the basis of sensory inputs:

> "[…] if one can identify a member of a given species and has some understanding of the innate releasing mechanisms characteristic of members of that species at the appropriate stage of development, then one will be able straightforwardly to predict what the creature will do when it registers stimuli of the appropriate type. Registering the relevant stimulus causes the appropriate response, and this can be fully

> understood, explained, and predicted without any appeal to an intermediary between stimulus and response. Similar input-output links can be seen in the case of sensorimotor schemas and various types of conditioned behaviour. Psychological explanations of behaviour only become necessary when no such input-output links can be identified." (Bermudez, 2003, p. 8)

Since behaviours listed at level 0 rationality *are* behaviours that can be predicted on the basis of sensory inputs alone, they don't require psychological explanations. If we are interested in understanding why we automatically move our fingers away from the fire when we get close to it, we don't need to employ an explanation that refers to our belief that if we keep our fingers on the fire they will burn, nor to our desire to move away from there. We can "simply" explain our behaviour in terms of instincts: we move away our fingers quickly and automatically because our movements are guided by our natural instincts. This is a simple example of a hard-wired behaviour (we know that, by observing another person getting close to the fire, that person will move away too) that is not peculiar to an organism, an animal or a person. A level 0 behaviour, therefore, doesn't result from a space of alternatives available to the organism in the here and now of its interaction with the environment. The possibility of other courses of action can be appreciated only from a phylogenetic perspective: by observing the species' evolution through time, we recognise that natural selection has provided the whole species with the foraging strategy that is adaptive for its survival in a certain environment.

With respect to point (ii), Bermudez argues that these behaviours are instances of rationality not only because they result from a range of possible other courses of action, but also because they match normative criteria. In particular, he claims, they maximise a certain kind of utility (e.g. the energy gained from food).

Utility is a quantity that depends on an organism's goal. This means that understanding utility (as the normative criterion necessary to specify the nature of a given behaviour) requires getting clear on the nature of its corresponding goal. I claim that the nature of utility in the case of level 0 behaviour makes it an inadequate benchmark to evaluate the rationality of behaviour.

Consider the case of an animal performing foraging behaviour. The animal's goal is a typical adaptive goal: surviving. Achieving this adaptive goal requires the exploitation of natural tendencies that all living systems have. There is no clear sense in which we could understand this goal (and its related utility for the animal) as related to a space of alternatives and the performed behaviour as the result of an individual choice. Let us call the utility that is related to shared natural goals "objective utility", and the utility that refers to individual specific goals (e.g. arrive at University as quickly as possible) "subjective utility".

I argue that the first kind of utility cannot constitute a normative benchmark against which the rationality of behaviour should be evaluated: a behaviour that maximises an "objective utility" can be explained in terms of adaptation and natural selection. The notion of rationality would be explanatory redundant. Indeed, foraging behaviours are listed among the classical examples of adaptive behaviours.

Addressing the presence of adaptive trait is, instead, insufficient to explain behaviours that maximise "subjective utilities". While natural selection can provide coarse-grained explanations of why animals have certain abilities, it can't explain why a specific animal chooses a *particular* action instead of another in a *particular* context. Natural selection can account for why we have the ability to make decisions, but it can't explain why it is the case that John chooses the shortest path to get to University. In other words, the notion of adaptation can explain and distinguish behaviours that are adaptive from those that are non-adaptive, but it doesn't seem suitable to explain how a cognitive system behaves in specific circumstances. The achievement of individual goals often requires the animal to select a behaviour on the basis of an expected utility maximisation that depends both on her goal and on her background knowledge: the richer the knowledge and expertise, the wider the space of alternatives the subject can choose from. The bottom line is that for a behaviour to be rational it is not sufficient that it matches some normative standards or that it is selected among a range of alternatives. It is first of all crucial to understand whether the standards are adequate benchmarks for comparison, and whether the space of alternatives is ontogenetically or phylogenetically determined. Given that external criteria are orthogonal to internal ones, I suggest that we are entitled to ask whether an adaptive behaviour is also a rational behaviour when:

- The goal of an animal is not straightforwardly linked to survival or reproduction
- The behaviour matches some normative standard
- The utility that gets maximised depends on the animal's goal and knowledge, hence it is subjective and not objective
- The behaviour is one of a range of other possible behaviours available to the organism in a given context

Drawing on the claims I have just made, my answer to Bermudez's question: "are psychological explanations available all the way down the ladder of rationality, or is there a privileged level or levels of rationality below which psychological explanations is not possible?" (*ibid.*, p. 128) is twofold.

First, I believe that the question already implies that it is possible to have rationality even when we are not dealing with behaviours that require a psychological explanation. As I have just argued, this assumption is misleading: if a behaviour does not require a psychological explanation and if it can be predicted solely on the basis of sensory inputs, then that behaviour is not a candidate of rational explanation. Rationality can enter the picture only when we are dealing with behaviours that cannot be predicted solely on the basis of information about the current environment. I have suggested elsewhere that explaining these behaviours often requires the employment of internal representations too (see chapters 2 and 3).

Second, I have provided arguments in favour of the idea that rationality cannot be applied to level 0 automatic, hard-wired, stimulus-response behaviours because they can be accounted for in terms of adaptive traits.

5.4.2 – Level 1 and level 2 rational behaviours: where is the difference?

In contrast to level 0 behaviours, level 1 behaviours do require a psychological explanation and the space of alternatives is attributable to specific organisms. Level 1 behaviours are:

- Selected among a set of other possible behaviours
- Closely dependent on the here and now of the organism in interaction with its environment
- The result of selection processes
- Selected on the basis of direct perception because the content of perception already contains information about possible courses of actions (the animal perceives affordances)

In particular:

- Selection among affordances is not decision among affordances
- Selection among affordances is rational because it allows the animal to behave in a way that maximises a certain expected utility

A case of level 1 rational behaviour is, as I have previously analysed, that of perception: it can't be predicted from knowledge of sensory inputs alone because it results from a (limited) number of alternatives.

According to Bermudez, a behaviour is rational when it matches a normative standard and results from a range of alternatives, and a behaviour belongs to level 1 domain when its underlying reasoning process traffics in representations of actions or affordances. When an organism perceives a scene, for instance, its perception is mediated by internal representations (i.e. internal instrumental components), which have a very simple content that already dictates a particular course of action.

The nature of the underlying process is central. Bermudez claims that, through a selection process that depends on the content of affordances, an animal chooses which action should be performed to maximise the expected utility. Selection processes need to be kept distinct from decision processes because only the latter require the employment of beliefs.

With reference to the case of the animal confronting another animal, Bermudez claims that the choice between the two possible behaviours fight or flee "might […] be understood at a purely perceptual level. It is perfectly possible, and indeed highly likely, that the choice between such action-representations can be made on relatively simple and more-or-less noncognitive grounds" (*ibid.*, p. 123).

My main concern here is on the nature of the processes underlying level 1 behaviours. Bermudez claims that they result from selection and not from decision processes because they involve quite simple representations and do not necessitate the employment of any instrumental beliefs. To address this point, let me first summarise the main features of level 2 rational behaviours.

The central features of level 2 rational behaviours are the followings:

- Behaviours that result from genuine decisions
- Behaviours whose selection is made from a range of possible alternatives
- Behaviours whose selection is made on reasons that are consequence-sensitive
- Reasons are called instrumental beliefs and they are representations of contingencies holding between an action and its expected outcome
- The comparison of representations of contingencies is not inferential, but immediate and straightforward
- These behaviours happen when organisms cannot directly perceive their goals in the environment or the courses of action that would yield the desired results
- Operational criteria show that behaviours are level 2 rational behaviours when organisms change them with regard to changes in the contingencies between actions and desired results

Bermudez affirms that a genuine decision process does not need to confirm to Fodor's step-wise model; rather, it needs to be made on consequence-sensitive grounds: an animal decides to perform an action because it has predicted that that action will lead to a greater utility (with respect to its goal and to the other possible actions). This means that the selection of an action in cases of genuine decisions depends on an organism's ability to predict the consequences that different courses of action will have in a certain environment.

Making predictions, forming expectations and deciding in an anticipatory manner are peculiar features of organisms performing level 2 rational behaviours. Level 1 rational behaviours, on the other hand, do not involve instrumental beliefs because organisms immediately perceive how to satisfy their goals.

My first critique to the distinction between level 1 and level 2 rational behaviours concerns Bermudez's assumption that affordances are *not* about contingencies between actions and their possible consequences in a given environment. Why does Bermudez claim that selection processes operating on affordances are made on non-consequence-sensitive grounds? Why can't the distinction between the reasoning processes implicated in level 1 and level 2 rational behaviours be understood in terms of degree rather than kind? Indeed, selection among affordances seems to depend highly on the likely consequences of the different available behaviours. If the selection in a level 1 behaviour is goal-oriented, then it is also consequence-sensitive in the sense that the resulting behaviour will be the one that is expected to accomplish the goal. I therefore suggest that a level 1 behaviour results from a selection that is sensitive to the goal of the animal and to the match between its expected consequences and the achievement of the goal. This process has to be explained in consequence-sensitive terms. So, where might the difference lie?

Working within the Reinforcement Learning (RL) framework[24], Daw and colleagues (2005) suggested that there are different neural mechanisms that contribute to action selection: model-based and model-free mechanisms. An animal working with a model-based approach can rely on its internal knowledge about the causal structure of the world to construct predictions on the fly of the long-term outcomes of various actions. This strategy, which can be computationally expensive in terms of the amount of memory and time required, allows the animal to react to changes in the environment in a straightforward way. A model-free strategy, instead, enables the animal to adjust to new contingencies only after it has acquired enough experience with the new environment. The difference between model-based and model-free strategies is similar to Bermudez's difference between level 1 and level 2 rational behaviours.

Daw and colleagues ran a series of experiments on rats trained to press a lever to obtain food. They found that they used different RL strategies in different circumstances: when animals were moderately trained on the task, their decisions were sensitive to outcome devaluation (i.e. reward value of food was reduced by either feeding the rats or by making the food poisoned), while their choice was insensitive to changes in contingencies when they were well trained on the task. Interestingly, the adoption of one of the two strategies varied on the basis of how time-demanding and complex the task was and on how trained the rats were on that task.

Generalising these empirical results, we can say that if an animal needs to choose an action in a time-demanding situation without previous extensive experience with that situation, then its selection will be based on instincts. In this

[24] The Reinforcement Learning (RL) framework offers models of optimal and approximately-optimal learning of which behaviour achieves a goal in face of uncertainty or rewards. In RL, the decision of which action to undertake is based on each of the possible available actions' predicted values, which are defined in terms of the amount of reward that each action is expected to bring (see chapter 6 for more details on RL models).

case the animal's behaviour is a level 0 non-rational behaviour. If, instead, an animal is not trained on the task but has time to select the action, as it is in the case of the rats in the example above, then it will be sensitive to changes in the contingencies (i.e. it will perform a level 2 (non-habitual) rational behaviour).

Other possible scenarios are those where an animal has been extensively trained in a given task. Here the complexity of the task is crucial. Daw and colleagues showed that if rats were adequately trained on a task where they didn't have a wide range of possible actions to reach their goal (i.e. the task was quite simple), they remained insensitive to changes in contingencies until they acquired more experience with the new settings. Since they didn't have a set of adequate habitual behaviours to bear on the situation, rats needed further training to learn that other responses to the environment were adequate with respect to their goal. If, instead, they experienced various possible courses of actions during their training period, all of which were adequate with respect to their goal, rats would be sensitive to devaluation.

Interestingly, level 1 behaviours do not correspond to any of the above scenarios: animals perform on the basis of model-based or model-free strategies in different circumstances and in relation to their degree of knowledge of those circumstances. The processes that are responsible for their behaviours are not different in kind; rather, they are different in the amount of time available, in the level of complexity and in the richness of the animals' background knowledge. Contingencies between actions and their expected outcomes are in place in all the different situations, except when rats have no time to adequately consider the situation and no previous experience with the task. If operating on consequence-sensitive grounds is what is required, in Bermudez's account, for a behaviour to be a level 2 behaviour, then there seems to be no room left for level 1 rational behaviours.

Consider the following example. Imagine we plan to go to a park populated by many wild animals, some of which are bears. We decide to go with a friend who is familiar with the park. He tells us that we might encounter a bear and that, if this happens, we shouldn't be afraid: the bears living in the park are innocuous. We go and visit the park when, at a certain point, we turn around and find ourselves face to face with a bear. This is a situation where we clearly have little time to plan our next move. According to Bermudez, in this situation, we will select the action that, compared to others, is expected to maximise the current utility. We will probably have two alternative courses of action: confront the bear or run away. If our goal is to survive, we might select the running behaviour since it is the one that will maximise our expected utility. Our selection here, Bermudez would say, is not influenced by our background beliefs (e.g. our friend told us that the bears that live in the park are innocuous), but only by fear.

Consider now a slightly different scenario. We are in the park and we see a bear in the distance. We have some time to decide what to do before the bear gets too close to us. We might start thinking that, although we are scared and want to run away, our friend told us that these bears are not dangerous. Instead of selecting between the two affordances, we might want to consider the possibility of letting the bear getting closer to take a memorable picture.

If we now compare the two scenarios, we see that they are similar in terms of their contexts, but different in terms of their goals. While in the first scenario the bear is too close for us to engage in a process where our background beliefs can actually make a difference, in the second scenario we can recall what our friend told us (i.e. the bear is not dangerous). In this second scenario, then, we could decide to wait and take a picture instead of automatically reacting to the situation by running away. Bermudez would explain the first case as involving a behaviour resulting from a selection among affordances, which are not consequence-dependent nor need to involve beliefs. Only the second case would be for him an instance of selection, hence of level 2 behaviour. Nevertheless, if we analyse the two scenarios, we notice that the difference between level 1 and level 2 behaviours (i.e. the difference between instrumental components, instrumental beliefs and processes operating over them) is more blurred than the one presented in Bermudez's account.

In the next section I will address some reasons why Bermudez might have misunderstood the distinction between level 1 and level 2 rational behaviours.

5.4.2.1 – The methodological inadequacy of operational criteria

Bermudez believes that the evidence that proves the difference between selection and decision processes comes from experimental results. He claims that the analysis of experimental data with the use of some minimal operational criteria, as he calls them, can distinguish a behaviour that results from genuine decision processes, hence a level 2 rational behaviour, from a behaviour resulting from selection processes, which is, instead, an instance of level 1 rationality.

He affirms that only a behaviour that changes once external conditions are different results from decision processes over instrumental beliefs about the contingencies between actions and their consequences. On the contrary, a behaviour that doesn't change once external circumstances change is an instance of selection processes over instrumental components (but not instrumental beliefs). Interestingly, Bermudez claims that organisms operate over internal components in immediate and straightforward ways in both cases.

The point I want to make in this section is that external criteria are not sufficient to explain the nature of rational behaviour. In particular, I claim that behavioural criteria can guide the search for explanations of rational behaviour, but that additional information concerning the nature of the internal and underlying reasoning processes is necessarily to say whether a behaviour is similar or different to another one, whether two behaviours are of the same kind or not, and whether they are instances of rationality.

What operational criteria *can* show is that there are some differences in behavioural outcomes and that these differences *might* be due to distinct processes: one kind of process that operates on simple forms of representations and on a restricted set of alternatives (i.e. selection processes), and another type of process that requires the comparison of more complex and numerous representations (i.e. decision process). Bermudez claims that the second kind of process happens when the goal is not immediately perceivable by the organism or when the organism does not immediately know which is the most appropriate action.

Nevertheless, we could explain these results without invoking the existence of two different underlying reasoning processes, but only a difference in terms of their complexity and in terms of the components used. In the same way in which Bermudez himself justifies the rational nature of a level 1 behaviour even if the range of alternatives is limited and the nature of the representations employed is simple, we could hypothesise that the difference between a level 1 and a level 2 behaviour is in terms of the number of actions available to an organism and of the complexity of the representations involved. Accordingly, they would not involve two distinct kinds of processes. As I argued in the previous section, it could be the case that when an agent needs to behave in a very fast way in reaction to the environment, her behaviour is more dependent on instincts than on background beliefs. If this were the case, then, the behaviour would simply be an instance of level 0 rationality: it would result from a hard-wired mechanism that natural selection has offered us to cope with the environment.

5.4.3 – Only a terminological disagreement?

Someone might think that my point of contention is purely terminological: since Bermudez has clarified the notion of rationality he employs in his analysis and since he is consistent with his definition, we can agree with his conclusions too.

I want to resist the idea that my criticisms of Bermudez’s analysis of the notion of rationality are purely terminological.

First, I believe that to make progress in our understanding of rationality we need to agree *at least* on some minimal criteria for a behaviour to be rational. Having a space of alternatives and matching some normative standards, however, are too general criteria for rational behaviour. We could, for instance, end up listing as rational a behaviour for which the notion of adaptation can do the explanatory work already. As I claimed in 5.4.1, applying the notion of rationality to hard-wired and instinctual behaviours does not offer any explanatory gain over and above that provided by evolutionary explanations in terms of adaptive processes.

I also believe that a clearer idea of what rationality might amounts to is necessary to avoid certain ambiguous conclusions about the nature of cognitive behaviour. Consider the conclusions that Bermudez draws from the idea that the reasoning processes involved in both level 1 and level 2 rational behaviours are not inferential in nature. Starting with this premise, Bermudez concludes that rationality does not have to be considered in terms of the ability to constructs arguments or in terms of the ability to employ strategies and inferences prescribed by rational theories. According to him, a behaviour is optimal because it can be modelled in terms of expected utility theory (and it can be compared to standards) even if it doesn't result from its application. The foraging behaviour mentioned above is such an example: an animal behaves in a way that maximises the energy gain from food not because it has previously and internally calculated which action would have been optimal given its goal and the current environment, but because it follows simple (and innate) rules and heuristics. An example of a level 1 behaviour that is externally rational because it can be modelled according to a theory of expected utility, but whose outcome does not result from the application of such a theory, is the case of the animal that is confronting a dangerous animal. Here, Bermudez claims, the selection of one of the two courses of action, fight or flee, is based on direct perception, that is, on a simple heuristic that allows the behaviour to be fast and also appropriate. The animal does not calculate utilities. All the relevant information is already present in the context of perception. In these cases, appropriateness is understood as a matter of an animal *straightforwardly* acting on a given instrumental component or belief.

From his analysis of the reasoning processes implicated in the different rational behaviours, he concludes not only that there are behaviours which are rational and, at the same time, non-cognitive or not based on decision processes, but also that we are able to apply models and theories to assess the rationality of a behaviour even though that behaviour does not result from the application of those models and theories *at all*. Rationality, Bermudez argues, can be achieved even if the behaviour in question does not conform to the strategies prescribed by classical rational theories. He then goes on to argue that rational calculations are too complex and demanding given humans' cognitive limitations and that experimental data show how, in many reasoning situations, human performance deviates from that expected on the basis of classical rational theories. Humans, Bermudez (2000) claims, do not follow strategies and inferences prescribed by such theories. The reasoning processes used in these tasks are not imperfect applications or approximations of techniques prescribed by rational theories because they are not their application *at all*.

This claim is in line with the proposal according to which our rationality is bounded. Advocates of bounded rationality or of the so-called "ecological standards of rationality" (e.g. Gigerenzer *et al.,* 1996, 1999, 2000; Evans & Over, 1996, 1997) offer an account of reasoning and rationality in terms of heuristics and biases. In contrast to classical rational analysis according to which rational explanations consist in the specification of a goal and the environmental and in the subject's ability to derive an optimal solution to achieve that goal in that environment (Anderson, 1990), advocates of bounded rationality believe that cognitive systems are limited and unable to derive such optimal solutions; rather, they perform in ecologically successful ways by relying on fast and frugal heuristics. By stressing the effects of natural cognitive limitations on classical theories of rationality and reasoning, they argue that concepts of logic, decision-theory and probability do not match naturally with the strategies that humans employ in everyday reasoning tasks. The role of inferences should be assessed in terms of their ecological success in dealing with the world.[25]

Bermudez's proposal seems to be in line with these ecological accounts. He claims that once we pay attention to the nature of the underlying reasoning processes, we recognise that behavioural success can be achieved by employing strategies that are different from those prescribed by classical rational theories. Which are the data Bermudez relies on to draw this conclusion?

As I previously discussed, Bermudez argues that rationality is a matter of heuristics rather than inferential processes by relying on operational criteria: if the animal chooses differently once the circumstances are changed, then the animal can recognise and internally represent the presence or absence of specific contingencies between actions and their outcomes. Nevertheless, as I have argued above, relying only on operational criteria is a too weak strategy to justify the distinction between level 1 and level 2 rational behaviours. If my argument there was sound, then, the central distinction between affordances and instrumental beliefs and that between selection and decision processes can be put into question.

I have shown some possible consequences of drawing conclusions about the nature of behaviours and about their underlying processes before we actually have enough relevant information about them. We might end up claiming that all rational behaviours result from strategies that do not have anything in common with those prescribed by classical rational theories or that we have explained a behaviour once we are able to predict it. It is therefore important to recognise that explanations that

[25] However, consider the distinction between rational calculation and rational description (see Chater *et al.*, 2003). In contrast to what advocates of ecological rationality argue, it is plausible to claim that classical rational theories do have an important explanatory role, but only as normative criteria. This would mean that, given that the complexity of optimal calculations would exceed cognitive systems' real capacities, rational theories do not provide information about the computations actually carried out by cognitive systems. The real processes that operate in real-life reasoning processes only approximate those predicted by rational theories. Nevertheless, saying that classical theories can only be approximated does not deny their usefulness in accounting for rational behaviours.

derive only from external criteria cannot be considered adequate explanations. By adopting them we run the risk of interpreting behaviours in a way that is too dependent on our *a priori* commitments.

Accordingly, Bermudez's three levels of rationality can't be taken for granted and the explanatory power of purely personal-level stories about rationality and reasoning can be put into question. As seen in the previous chapter, autonomy theorists believe that personal-level folk-psychological explanations are necessary and sufficient for explaining and predicting others' people behaviours as they approximate to an ideal of rationality. Their hypothesis is that we can understand and predict others' behaviours to the extent that we can understand the inferential relations between their mental states and their actions (e.g. Davidson, 1980). According to autonomy theorists, only personal-level explanations can properly account for human rationality and we do not need to know much about the particular underlying machinery of such capacity. Once again, the evidence here comes from external observational criteria.

As claimed in chapters 1 and 4, good explanations of rational behaviours can't proceed only at the personal or external observational level. Operational criteria and rational expectations are not sufficient for explanation. What makes a process a reasoning process can't be simply derived from behavioural outcome. Rather, we need to complement the external perspective with an internal study of the processes in question. If there are indeed different components and if these different components are actually employed in different processes, then these differences should map onto the structure that makes them possible. Only once these differences are shown not only at the functional but also at the structural level, we can conclude that they actually exist.

Mechanisms are, therefore, necessary: a complete explanation of a capacity in terms of functional properties requires the identification of structures that possess these properties. This means that the distinction between instrumental components, beliefs and desires — if real — has to constrain the mechanism in the sense that distinctions between these representations and distinctions between the ways in which they are processed should be found at the level of the brain. At the same time, the identified mechanism should constrain the functional analysis of the capacity: the representational formats and the operations on them should vary depending on how the mechanism handles their distinctions.

Accordingly, whether a system implements a certain process instead of another one and whether it contains instrumental components or instrumental beliefs is also a matter of its structural features. Information about structural components, their organisation and their activities have to be found at the subpersonal-level of analysis. It is only once we recognise that subpersonal-level states and events are important to understand rationality constraints that we start asking about those specific mechanisms that underlie the behaviours that we call, from a normative perspective, rational. Given that Bermudez's three levels are characterised by the same rational standard — the maximisation of a certain kind of utility — what distinguishes them is the way in which "appropriateness" is calculated and achieved. Although Bermudez recognises this important distinction, the methodology he adopts is inadequate because it remains at a purely functional and behavioural level: he identifies internal components on the basis of their functions within the system and with respect to certain behaviours. He does not dig deeper into the structural nature of these components. Whether affordances and instrumental beliefs are the responsible components of different processes and whether they are actually processed in different ways can be ultimately proven only once we have information about the structural components that implement these supposedly different functions. Indeed, we already have some insights about rational processes in humans that do seem to follow the strategies prescribed by rational theories, and, in particular, Bayesian rational theories (see chapters 6 and 7).

5.5 – Conclusion

To summarise, my main concerns with Bermudez's accounts of rationality are the followings:

- Level 0 rationality does not identify rational but only adaptive behaviours
- The very distinction between level 1 and level 2 rationality is unclear:
    - Affordances, as well as instrumental beliefs, seem to be consequence-sensitive
    - Bermudez justifies the distinction between affordances-based processes and instrumental beliefs-based processes solely on the basis of operational criteria
- Operational criteria are not methodologically strong enough to establish the nature of reasoning processes since they consider only behavioural outcomes and how these outcomes are affected by environmental changes
- Conceiving the nature of reasoning processes that yield rational behaviours in terms of fast and frugal heuristics on the basis of operational criteria is unwarranted
- Claiming that in real-life reasoning situations humans do not follow strategies and inferences prescribed by rational theories because this is what some experimental results show is not enough to conclude that humans do not follow these strategies *at all*
- A functional analysis of the capacity to perform in context-appropriate ways needs to be supplemented and constrained by information about the subpersonal states and events responsible for that capacity
- Personal-level explanations cannot stand on their own. In order to offer an adequate explanation of a capacity, a mechanistic story is necessary too.

For all these reasons, I think that Bermudez's conclusions are too quick. I showed that he doesn't have enough information to support the idea that human rationality is only a matter of heuristics and biases. I argued that information about subpersonal mechanisms are required to genuinely explain why humans (and other animals) are often context-appropriate in their responses to the environment.

In the next chapter I will examine more closely the explanatory gain that derives from taking on board subpersonal information in the explanation of behaviour.

Chapter 6 - The Bayesian Neurocomputational Framework

6.1 – Introduction

So far I have analysed three different frameworks adopted to explain cognitive behaviour: the folk-psychological, the anti-representational and the physiological subpersonal frameworks.

With respect to folk-psychological explanations, I argued that: (i) their main goal consists in making behaviours intelligible rather than in explaining them; (ii) they appeal to an unclear notion of cause.

Regarding anti-representational explanations, I highlighted that they are more suitable to describe how a system's behaviour changes over time rather than to explain how a system performs a certain behaviour in the first place. This results in mathematical formalisations of a system's performance that bear little or no biological plausibility. Indeed, such plausibility, I argued, depends on the identification of localised mechanisms, which is not the focus of anti-representational explanations.

Concerning physiological subpersonal explanations, I argued that they cannot adequately explain cognitive behaviour because certain personal-level notions are required to explain cognitive phenomena and because neuroscientific practice doesn't provide evidence for the possibility of reductions of mental phenomena to molecular phenomena.

The goal of the current chapter is to examine the neurocomputational framework by analysing a family of models — the Bayesian models — that belong to it and their application to the study of different cognitive phenomena.

6.2 – The neurocomputational framework

Neurocomputationalism is a framework of explanation adopted to study human behaviour that considers the brain as a processor of information and cognition as a result of neural computations: humans perform cognitive tasks because they process internal representations in certain ways.

Neurocomputational explanations are special kinds of subpersonal explanations that try to explain why humans perform cognitively in light of the various ways in which the nervous system copes with its environment. This makes the neurocomputational methodology different from the standard procedures in psychology and cognitive psychology: rather than studying behaviour independently of knowledge of the brain[26], the neurocomputational framework aims at uncovering the internal structure of the processes underlying various cognitive performances and, in particular, the internal biological components responsible for certain behavioural outcomes. Neurocomputationalism tries to achieve these results by suggesting mathematical models and by relying on simulations. It first identifies the task a subject needs to perform and it then suggests a possible way in which she can

[26] In standard psychology and cognitive psychology, experiments are designed to measure and identify only behavioural information (i.e. reaction times, patterns of errors, and so on). This information is then used to understand the nature of the underlying cognitive processes.

carry it out. This way is defined in terms of algorithms that, if executed, should yield the desired behavioural response.

The neurocomputational methodology follows the basic assumption that if a system has a certain property, then that property is not a basic fact about the system, but depends on the nature and organisation of its components parts (e.g. Glennan, 2010). This methodology has benefited from the rise of new techniques for the study of the brain (e.g. fMRI and PET), which has allowed researchers to envisage a future where both functional and structural features (e.g. specific brain areas, neurons, populations of neurons, neurotransmitters, synaptic connections, and so on) of cognitive processes could be uncovered. The hope is that these (and other) techniques will help the identification of possible correspondences between stages of information processing responsible for a specific cognitive phenomenon and repeatable events at the level of the brain.

At present, only tentative bridges between neurocomputational models and brain processes have been suggested with respect to quite simple perceptual and motor tasks, but researchers expect that more bridges will be discovered in the future.

The novelty of the neurocomputational framework can be better understood within the three-level framework laid out by David Marr (1982). Marr's framework specifies the computational task that the system needs to solve together with a class of rules or algorithms that can be responsible for the system's success in the task. It also suggests ways to uncover the implementational nature of the cognitive process by relating informational stages to biological transactions among neurons and population of neurons. As I will show in the current chapter and, especially, in the next chapter, the simultaneous focus on computational, algorithmic and implementational levels makes the neurocomputational framework apt for co-evolution. This means that information at different levels can interact and constrain each other, eventually leading to modifications of some of the categories that we normally adopt to explain and understand cognition and behaviour. In the words of

Patricia Churchland (1986), co-evolution is a fruitful framework within which our understanding of brain and mind “may need to be revised, and the revisionary rationales may come from research at any level” (*ibid.*, p. 746).

Neurocomputational explanations make reference to brain components and to the transformation of information among neural populations. They suggest possible ways in which neural processes might encode, use and transform information in terms of neural computations (i.e. transformations of neural spike trains according to algorithms). It is by describing how neural components encode information and how they interact with other components to transform this information that neurocomputational explanations aim at accounting for how cognitive functions and behaviours are generated within a cognitive system.

In the rest of the chapter I will examine the neurocomputational framework mostly by analysing one of its families of models – Bayesian models – and their way of accounting for various cognitive phenomena.

6.3 – The Bayesian neurocomputational framework

We access the world through our senses, which are our main sources of information. Sensory information, even though vital for successful interactions with the world, is often noisy and ambiguous. Not only is sensory information ambiguous, the world often presents itself in quite uncertain modes too. The same object seen from different perspectives, for instance, will yield different sensory information and the same sensory information can be caused by different environmental states.

How can we overcome these ambiguities and extract information about the state that obtains in the world? How can we select actions appropriate to the current circumstances and to our goals on the basis of noisy and uncertain data?

The task of the brain might seem impossible at first: it must infer information about the likely causes of its sensory inputs without any direct access to them. All that the brain "knows" is the way in which its own states (e.g. spikes of neurons) flow and modify.

The Bayesian neurocomputational framework suggests a possible way in which this uncertainty can be handled, thus allowing agents to behave successfully in the world. The main hypothesis at the heart of Bayesianism is that cognitive agents use some rule (or approximate rule) when perceiving the environment, making decisions and performing actions. In particular, the claim is that the main goal of our nervous system is to infer the cause of its sensory inputs by relying on ambiguous and noisy sensory information and internal generative probabilistic models of the relevant variables in the environment causing the sensory stimuli. Internal probabilistic models are then tuned by learning and experience via the interactions of the agent with the environment.

The process of making informed guesses about the causes of sensory stimulation and updating those guesses based on new evidence is called Bayesian inference, and it results from the execution of Bayes' rule:

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

Bayes' rule indicates a way in which the nervous system can update the (posterior) probability [P(H|S)] that a certain hypothesis (H) is true given the sensory data (S). Accordingly, Bayesian inference is possible by multiplying the probability of the sensory data given the hypothesis (likelihood) with the probability of the hypothesis (prior), divided by the probability of the data.

Advocates of the Bayesian neurocomputational framework believe that the cognitive system combines uncertain information about stimuli with prior information in a way that accounts for their uncertainty. This means that the internal representations over which Bayesian computations are performed need to encode both the value and the uncertainty of the stimuli. The estimation of the uncertainty of the sensory stimuli modifies the weighting of prediction-error, which, in turn, impacts the higher levels of the hierarchy (see 6.5). In other words, generative models must include what Jacob Hohwy calls "precision expectations": context-based assessments of the reliability and salience of the sensory information itself. Precision-expectations allow the system to require precise prediction-error signals on some occasions and less precise prediction-error signals on other occasions. In Hohwy's words:

> "Prediction error that is unreliable due to varying levels of noise in the states of the world is not a learning signal that will facilitate confident veridical revision of generative models or make it likely that selective sampling of the world is efficient. Prediction error minimization must therefore take variability in prediction error messaging into consideration – it needs to assess the precision of the prediction error." (Hohwy, 2012, p. 4)

According to this framework, our cognitive system can be described as a combination of top-down and bottom-up signals. Top-down signals are the prior expectations about the state of the world before the system receives sensory information [P(H)], while bottom-up signals are the sensory information conditional on the priors [P(S|H)].

Priors heavily influence what agents perceive and represent by constraining the way in which sensory information is processed. Predictions about current environmental cues are then selected according to the best model the system has of the possible causes.

Learning is driven by mismatches between bottom-up and top-down signals: if there is a mismatch between the prior expectation and the bottom-up sensory signal, the prior gets corrected using Bayes' rule within a cascade of cortical processes where the higher levels attempt to predict the input at the lower levels on the basis of their current model of the causal structure of the environment.

The signal of the mismatch is called prediction-error. Changing the priors via prediction-errors coincides with changing the predictions about the causes of the lower-level activities.

To get clear on Bayesianism, and on the neurocomputational framework more in general, consider the case of perception.

It is now widely accepted that perception takes place along a cascade of many processing stages over cortical areas arranged in a hierarchical structure. According to the Bayesian neurocomputational framework, neural activations produced by sensory inputs from the current visual scene belong to the lowest level of the hierarchy. These inputs are processed along a cascade of multi-level stages, where each level tries to predict the activity at the level below it via top-down (backward) connections. Top-down influences allow the activity at one stage to become input to a lower-level stage. As long as top-down signals successfully predict the lower-level activity, no update or modification is required. If, instead, the top-down predictions don't match the activity at the lower level, a prediction-error signalling the inappropriateness of the prediction with respect to the current activity is generated and propagated to the higher levels. This is the situation where the hierarchical model needs to be updated: probabilistic representations at the higher levels are modified to allow the next top-down predictions to cancel the prediction-error and to yield an appropriate perceptual inference.

An example of these models at work is Rao and Ballard's model of visual processes (1999). The visual cortex is here characterised in terms of top-down and bottom-up signals. In particular, it has been suggested that area V2 in the visual cortex might send top-down predictions of the activity in V1, and that V1 might send bottom-up error signals to V2.

These ideas have been tested on a hierarchical neural network, showing that it could successfully predict the external causes of the sensory inputs by computing algorithms that reduced prediction-errors within a cascade of processing stages. The network also showed that each level in the hierarchy could deal with different features of the cause: lower-level stages could encode simple features of the stimulus (e.g. the object's orientation), and higher levels could encode more general and abstract features of it (e.g. the object's larger spatial configuration).

In this network, at the lowest level, there is some pattern of energetic stimulation produced by the patterns of light in the current visual scene. These signals are then processed via a multilevel cascade where each level attempts to predict the activity of the level below. The predictions allow the activity at one level to become the input to the level below. As long as the predictions are successful, no further action is required. If, instead, there is prediction-error, the error-indicating activity gets propagated higher up in the hierarchy. This process adjusts the probabilistic representations at the higher level so that the following predictions can cancel the prediction-error at the lower level. As Rao and Ballard put it:

> "The prediction and error-correction cycles occur concurrently throughout the hierarchy, so top-down information influences lower-level estimates, and bottom-up information influences higher-level estimates of the input signal. Lower levels operate on smaller spatial (and possibly temporal) time scales, whereas higher levels estimate signal properties at larger time scales because a higher-level module predicts and estimates the responses of several lower-level modules […]." (Rao & Ballard, 1999, p. 80)

Rao and Ballard specifically worked on a three-level hierarchical network of predictive estimators, which was trained on image patches extracted from five natural images (see Figure 1).

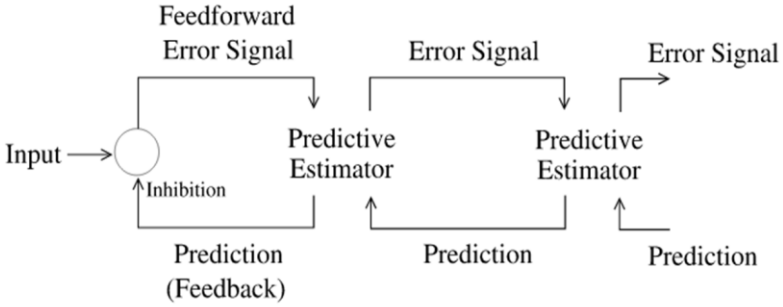

Figure 1: General architecture of the hierarchical predictive coding model. At each hierarchical level, feedback pathways carry predictions of neural activity at the lower level, whereas feedforward pathways carry errors between the predictions and actual neural activity. These errors are used by the predictive estimator (PE) at each level to correct its current estimate of the input signal and generate the next prediction. (Rao & Ballard, 1999, p. 80)

Using learning algorithms that progressively reduced the prediction-error across the cascade of processes, the network successfully learned to use the responses of the first level to infer features of the natural scenes, such as oriented bars and edges. The second level, instead, learned to capture the various combinations of features represented at the lower level. These combinations corresponded to patterns involving larger spatial configurations (see Figure 2). This means that the network was able to construct a generative model of the structure of the sensory information by relying only on the statistical properties of such information.

These were early and relatively low-level results, but the learning model itself has proven useful and highly applicable. Further studies (Knill *et al.*, 1996; Knill & Pouget, 2004; Friston & Stephan, 2007) have shown that people seem to perform in a manner similar to Bayesian models in a wide range of perceptual and motor tasks.

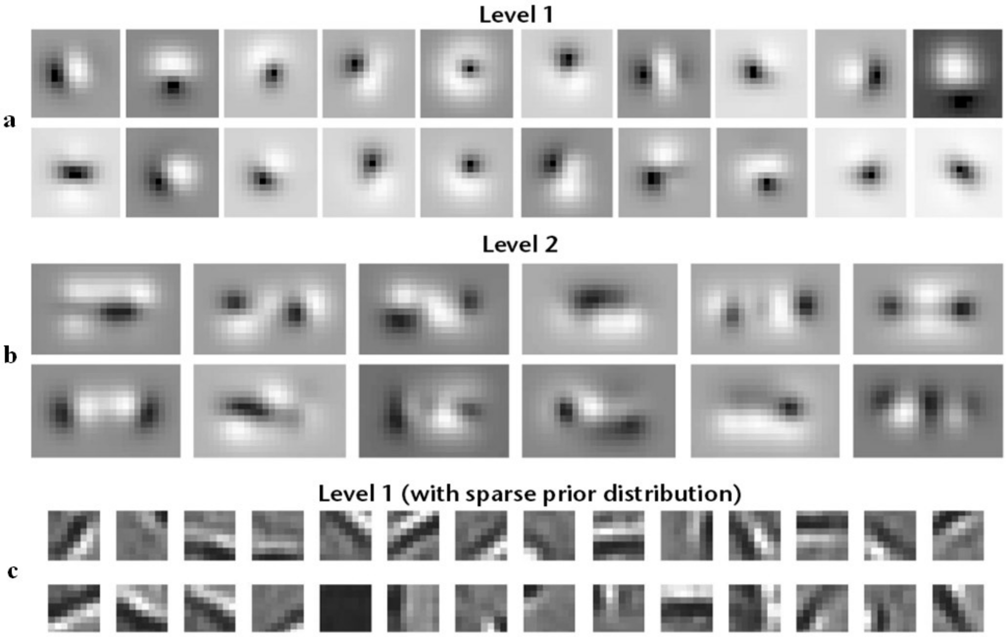

Figure 2: (a) First bottom level that represents basic features of natural scenes, such as oriented bars and edges. **(b)** Second higher level that represents larger spatial configurations of features of the stimuli. **(c)** Another image of the representing activity of level 1.

6.3.1 – Binocular rivalry

Hohwy and colleagues' hierarchical predictive coding model of binocular rivalry (2008) is a particularly good example that doesn't restrict to very low-level visual phenomena.

Binocular rivalry is a striking visual phenomenon that occurs in experimental conditions when the eyes are presented with two different images and the resulting subjective perception alternates between the two. The right eye might, for instance, be presented with the image of a house, while the left eye receives an image of a face. Under these conditions, subjective experience is "bi-stable", that is, it alternates between the house and the face.

Although there have been many studies on the nature of this phenomenon, the responsible mechanisms at play are not well understood. Hohwy et al. believe that adopting a different theoretical framework could help to make sense of apparently many different findings related to binocular rivalry, thus advancing our understanding of it. In particular, they argue, within this alternative framework the phenomenon becomes a reasonable response to an unusual stimulus condition.

The idea here is that the Hierarchical Bayesian framework can provide the computational mechanism that best explains binocular rivalry. Accordingly, the cognitive system tries to match top-down predictions with bottom-up signals, which are caused by the states of affairs in the environment. When the matching is good, the bottom-up signal is "explained away" (*ibid.*, p. 694). In this context, the best hypothesis is the one that makes the best prediction and that, ultimately, determines the content of the resulting perception. Other possible hypotheses are effectively inhibited. In the case of binocular rivalry, this means that top-down predictions explain away only the elements of the bottom-up signal that conform to the current best hypothesis; however, bottom-up signals contain information of both images (house and face). When one of the two images is selected as the best hypothesis, there remain certain elements of the driving signals that the current winning hypothesis doesn't predict. This results in an increase of prediction-error for the alternative hypothesis that gets propagated upward in the hierarchy. To suppress this prediction-error, the system needs to change hypothesis. But again, a large prediction-error signal emerges.

The persistent presence of prediction-error makes the perceptual inference unstable. This, in turn, causes perceptual alternation:

> "Alternation ensues in rivalry conditions specifically where there is a large unexplained but explainable error signal. In Bayesian terms, in this situation no one hypothesis has both high likelihood *and* high prior, and inference becomes unstable." (*ibid.*, p. 697)

The reason why we do not perceive a combined image is that we have certain *hyperpriors*, that is, "*a priori*, our brain has learnt that there can be only one cause of sensory input at the same place and time. This generic prior constraint reflects the way we sample the visual world; binocular vision, in primates, rests upon both eyes foveating the same part of visual space" (*ibid.*, p. 692).

6.3.2 – Cue integration

People behave in a manner similar to that predicted by Bayesian models in many different perceptual and motor tasks. However, little is known about the nature of probability distributions in the brain (e.g. how they are encoded, how they are transformed, how neural circuits can represent Bayesian inference, and so on).

Ma and colleagues (2006) aimed at uncovering something about the neural basis of Bayesian optimality by exploring the following two predictions:

- Neural circuits must represent probability distributions
- Neural circuits must be able to combine probabilistic representations in a nearly-optimal Bayesian way

The first challenge consisted in understanding how neurons could encode values of unique states of the world given their high variability. If neurons fire differently even in correspondence to the same state of the world, how can we understand which neuron encodes information about a specific environmental variable?

Ma et al. thought that one way to overcome this problem could be to study the firing profiles of population of neurons, rather than the activities of single neurons. If the stimuli are often noisy, then a plausible strategy that the brain might use to represent states of the world could be to encode the same stimulus' feature with several neurons. In this way, information could get encoded *redundantly* among neurons and this, in turn, could help to avoid problems due to, for instance, the loss of one specific neuron: if information about a state of affair in the world is encoded across multiple neurons, in case one of these neurons die or stops functioning normally, there won't be negative consequences for the overall functioning of the system. For this reason, researchers usually consider populations of neurons denoted via vectors of firing rates as loci of representation: population averaging helps reducing the network noise (e.g. Butts & Goldman, 2006) by relying on the cooperative encoding of multiple close neurons.

The first part of their study consisted in identifying the population activity (r) of a certain number (n) of neurons in a certain brain area (MT). This population's activity was then denoted by a vector $r^{MT} = \{r_1^{MT}, r_2^{MT}, \ldots, r_n^{MT}\}$. The spike count of a given neuron *i* in a certain time interval (δ) was formalised as $r_i^{MT}(t_n)$. When a stimulus (S) was presented, MT generated a series of patterns of activities that varied over time under the influence of neuronal variability. These patterns could be captured in terms of a probability distribution P(r|s). The optimal strategy for inferring the value of the stimulus from the probability distribution was to apply Bayes' rule to obtain the posterior probability distribution P(s|r).

Different methods can be used to extract one single estimated value from the posterior distribution. The most common method is the Maximum-Likelihood Estimation (MLE). This method allows the maximisation of the probability of the observed data given the distribution.

The typical way to assess the variability of spike trains is provided by the variance/mean ratio[27] of spike trains across trials. In the case of cortical neurons, such variability appeared to be described by Poisson-like statistics (i.e. the spike count's variance was proportional to the mean).

Ma and colleagues found that this kind of variability allowed neurons to represent probability distributions in a way that reduced optimal Bayesian inference to simple linear combination of neural activities.

The target of their study was cue combination, a perceptual task where the system needs to infer the value of a stimulus (S = spatial location of an object) from

[27] The variance/mean ratio indicates how sparse the population's activity is across a certain amount of time in relation to the area close to its pick.

noisy and ambiguous visual cues (C_1) and auditory cues (C_2).

According to Bayes' rule, the posterior probability (i.e. the location of the object) can be found by performing the following computation:

$$P(S|\, C_1, C_2) = P\ (C_1|S)P(C_2|S)P(S)$$

Ma et al. found that, when the prior was flat (i.e. when the subject didn't have prior expectations about the nature of the stimulus), the sum of the two Gaussian probability distributions $P(C_1|S)$ and $P(C_2|S)$, with variances and means proportional to each other, was equivalent to Bayesian inference. This means that neurons with higher firing rates (i.e. neurons that produced more spikes in a given amount of time) had lower levels of noise.

They then wondered how, given this particular kind of noise observed in the cortex, optimal inference could be achieved. Their suggestion was that Poisson-like variability allowed the computation of optimal inference in a particularly easy way: optimal Bayesian inference was reduced to linear combination of neural activities ($r_3 = r_1 + r_2$).

Ma and colleagues' study generated a number of predictions about neural activities and about behavioural performances, both of which could be tested via further experiments. One of these predictions was the following: if people behave in a Bayesian way in cue integration and their neural activity is Poisson-like, then the inference will result from a linear combination of neural activities.

The security of these predictions depends on both psychophysical studies and studies at the neural level (e.g. specific neural circuits must exhibit Poisson-like variability). The more secure these predictions are, the more information we have about the mechanism responsible for cue integration.

This and similar experiments seem to suggest that working within a neurocomputational framework allow to test models, to generate predictions that become more and more secure once knowledge of the brain is incorporated in them and, ultimately, to bridge information-processing theories to activities at the level of the brain.

In the following section, I will consider another piece of neurocomputational machinery that is often used to explain why and how agents can choose actions appropriate to the circumstances and to their goals on the basis of probabilistic internal representations.

6.4 – Decision-making processes: model-based and model-free

As I have already showed in discussing some of the previous examples, prediction-error minimisation seems to be the main building block of a mechanism that allows agents to perceive their environment. In this section I will claim that (reward) prediction-error minimisation seems also to be relevant to account for how agents learn to predict the consequences of their behaviours so as to optimize them.

Prediction-error stands for a signal of the mismatch between the expected and the actual outcome of a given action that an agent can use to update her expectations so as to make future predictions more accurate.

The Reinforcement Learning (RL) framework offers models of optimal and approximately-optimal learning of which action to perform to achieve a goal in face of uncertainty and rewards. In RL, action-decision is based on each action's predicted value, which is defined in terms of the amount of reward that the action is expected to bring. These predictions pose statistical and computational problems when the reward depends on a sequence of actions and when early actions only cause deferred rewards.

RL models, which have been extensively used in neuroscience to understand possible computational roles of certain neural signals within brain activity (Daw *et al.*, 2005), come in two main families: model-free approach and model-based approach.

One class of model-free RL approaches is the temporal-difference learning (TD). This learning is assumed to be responsible for habitual behaviours. TD learning considers the flow of experience that cannot be easily divided into discrete steps because the predictive stimuli and the rewards happen at different points in time. In this case, an action is selected on the basis of its associated scalar summary of long-run feature values. The goal of the system consists in predicting all the future rewards that are expected to occur given the current and the previous stimuli it receives.

TD learning is a variant of the Rescorla-Wagner model of classical conditioning, where the system is said to make predictions about the reward of a given future event, then to observe the actual event and, in case the expectation doesn't match the observation, to update its knowledge to make future predictions more accurate.

The key computational quantity at the basis of this form of learning is reward prediction-error. The Rescorla-Wagner learning rule is the following:

$$V_{new} = V_{old} + \eta(\text{outcome - prediction})$$

$$V_{old} + \eta(R - V_{old})$$

In the formula, V_{old} is the prediction of the future reward that the system makes at a certain point in time, η is the learning rate of the action plan that determines the degree to which each experience affects the prediction for the future, R is the actual reward obtained, and V_{new} is the updated reward based on the reward prediction-error.

According to the Rescorla-Wagner model, the rule is applied at the end of each conditioning trial to all stimuli present in the trial. On the contrary, TD learning takes into consideration the continuous flow of experience: predictive stimuli and rewarding outcomes occur at different points in time. At each point, the goal of the system is to predict all future outcomes given the current and the previous stimuli.

This means that, if the system predicts an incorrect reward value at time *t*, it will have a temporal difference prediction-error at time $t + 1$ due to the fact that the immediate reward plus the future expected rewards will be higher or lower than those originally predicted.

The prediction-error δ at time $t + 1$ is formalised as:

$$\delta(t + 1) = \text{outcome } (t + 1) + \text{prediction } (t + 1) - \text{prediction } (t)$$

The system uses this TD prediction-error to update its original prediction made only on the basis of the stimulus at time *t*. This update allows the system to learn the correct value of that stimulus:

$$V(t)\text{new} = V(t)\text{old} + \eta \cdot \delta(t + 1)$$

or

$$V(t)\text{new} \ = V(t)\text{old} + \eta[\text{ outcome } (t + 1) + \text{prediction } (t + 1) - \text{prediction } (t)]$$

How can this model help us to explain the computations performed by the brain to decide which action to take in a particular situation?

Interestingly, neuroscientists (Houk *et al.*, 1995; Schultz *et al.*, 1997; Schultz, 2010) have identified certain neural substrates that appear to confirm to the TD mathematical model. In particular, they found that the majority of midbrain dopamine neurons (75–80%) in the substantia nigra pars compacta (SNc) and the ventral tegmental area (VTA) show a rather stereotyped phasic activation following unpredicted rewards, which indicates that they might encode a reward prediction-error. This burst is supposed to play a crucial role in learning.

An example of these studies runs as follows. By examining subjects performing on a reward-guided decision task, investigators extrapolate the values of the parameters to be fitted into the TD model. In particular, they estimate the learning rate η. The performance of the model is then compared to the subject's behaviour and to other rival models to evaluate its adequacy. Once parameters are chosen and set, the model starts making predictions about the reward prediction-error signals that the subject is expected to generate on each trial if she is operating on the basis of the TD model under consideration. The estimation of the amount of prediction-error signal in each trail is then compared to the variation in BOLD signal (i.e. the flow of oxygenated blood in the brain that can be measured via fMRI techniques) at each step of the task to establish whether there is a co-variation between the two. This methodology consistently finds reward prediction-error signals in the VTA and in areas of the striatum (Shea, *forthcoming*).

Dopamine neurons' response occurs irrespectively of the sensory modality and of the spatial position of the stimuli. Rather, their activation depends only on the reward's probability and magnitude such that:

- If a reward is better than predicted, dopamine neurons get activated (i.e positive prediction-error)
- If a predicted reward doesn't obtain or if it is worse than predicted, dopamine neurons are depressed (i.e. negative prediction-error)
- If a predicted reward obtains as expected, there is no response in dopamine neurons

Neuroscientists interpreted these results by saying that dopamine responses appear to resemble the teaching signal of efficient temporal-difference reinforcement learning models. The same reward obtained by the system at different times does not activate dopamine neurons. Only increasing rewards provide continuing reinforcement via the dopamine mechanism.

It is important to highlight, though, that not all forms of learning and decision processes, as modelled by RL, are dependent on dopamine. There is both behavioural and neural evidence for a multiplicity of mechanisms for decision-making, some of which don't seem to involve dopaminergic activity. Daw and colleagues (2005) have suggested that the dorsolateral striatum and its dopaminergic afferents could be the locus for model-free strategies, such as the TD learning, while the prefrontal cortex might implement model-based strategies, thereby supporting goal-directed behaviour.

In contrast to the model-free approach, the model-based RL method doesn't operate by representing the rewards associated to different possible actions. A system working with a model-based approach relies on a source of internal knowledge about the causal structure of the domain of action (i.e. the contingencies between actions and their possible outcomes) to construct real-time predictions of long-term outcomes. This is achieved by chaining together the predictions of the immediate consequences of each action and by using a particular desired outcome to flexibly determine the complex sequence of actions that are needed to achieve it. To do this, the system explores all possible future situations. The model-based strategy can be computationally expensive in terms of memory used and time required to do the searching.

Since predictions are made on the fly, the system can react to outcome devaluation in a more straightforward way than in the model-free approach: when contingencies change, predictions change too. A model-based method then results in more flexible, hence context-appropriate, behaviour, without needing extensive training.

Consider a classic experiment where hungry rats were trained to press a lever in order to get food in a food magazine. This sequence of actions was usually followed by a reward (i.e. food pellet). To successfully perform the task, rats needed to represent possible actions, the transitions between steps underlying possible actions and the reward that was available from an appropriate sequence of actions. In the next phase of the experiment, the reward value of the food pellet was reduced by, for instance, giving food to rats or by making the food poisoned. At this stage, rats were tested to see whether their behavioural choice would change in the presence of the devaluated reward outcome.

Experimenters hoped to understand whether rats behaved on the basis of a model-free or a model-based strategy. If they acted on the basis of a model-free strategy, they would continue to behave in the same way even after the food was devalued. If, instead, they operated on the basis of a model-based strategy, they would change their sequence of actions in accordance to the new reward.

Interestingly, Daw and colleagues found that animals performed on the basis of both strategies in different situations. When animals were only moderately trained on the task, their decisions were sensitive to outcome devaluation. When animals were, instead, extensively trained on a given task, the selection became insensitive to devaluation. These results suggested that animals could change their selection strategies from model-based to model-free on the basis of training and experience in a certain environment and with respect to a given task. Lesions to dopaminergic inputs to dorsolateral areas of the striatum seem to block the transfer from model-based to model-free strategies (Yin *et al.*, 2004; Faure *et al.*, 2005).

Daw and colleagues also found that the strategy employed depended on the complexity of the actions animals were supposed to perform and on the proximity of the actions to the rewards. In more complex tasks (e.g. animals were extensively trained but could perform various actions to get reward), they remained sensitive to devaluation.

Daw et al. concluded with a very interesting claim: given that animals switch between the two strategies on different circumstances and given that both strategies aim at rational goals, the two approaches are normatively similar. In some cases, the model-free strategy can more efficiently accomplish the goal with respect to the model-based one.

In contrast to the model-free approach, the putative mechanism for model-based strategies is not well understood. Some evidence seems to suggest the existence of distributed neural areas that might be implicated in it, such as the dorsomedial striatum, the prelimbic prefrontal cortex, the orbifrontal cortex, the medial prefrontal cortex and parts of the amygdala (Dayan & Niv, 2008).

In line with Shea (*forthcoming*) and consistent with the claims I have made in previous chapters, these and other studies don't show that the brain operates on prediction-error signals, represents rewards and uncertainties and computes over them with TD or model-based algorithms; rather, the results indicate that there are at least some quantities similar to prediction-error and to the expected values that are processed in the brain and that probably play an important role in generating decisions.[28] Accordingly, instead of concluding that there is an identity between the TD model and the brain processes, we can say that some features of the phenomenon are captured by the model, but that their exact relationship remains still largely unknown.

6.5 – Empirical evidence

There are accumulating data indicating that the cortical network might implement Bayesian inference (Doya *et al.,* 2007; Knill & Richards, 1996; Rao *et al.*, 2002). Specifically, there are three main sources of evidence: psychophysical findings, computational models and known structural features of sensory systems.

Psychophysical experiments, which have motivated the search for correlates of Bayesian algorithms in the first place, are the most telling evidence for Bayesian

[28] "The ubiquitous problem with imaging methods that the brain activity being recorded may just be a side effect of, rather than the constitutive basis of, the information processing which gives rise to the behaviour in question has been partly addressed by obtaining converging evidence from a variety of sources (neurophysiology, fMRI, EEG, TMS, etc.). A more important problem concerns the validity of model-based analysis of fMRI data […]. It is likely that a whole family of algorithmic models would show a reasonable match to the empirical data […]. It is hard to differentiate the particular temporal difference learning model that is used to account for trial-by-trial variations in neural activity from other reinforcement learning models in which prediction error signals play a role." (Shea, *forthcoming*)

models (e.g. Berniker & Kording, 2008; Kording *et al.*, 2007). The cue integration task I have considered above is one of these examples.

There are also a number of computational models that show how approximate Bayesian inference could be implemented in biological neural networks as well as some structural features of sensory systems that speak in favour of hierarchical Bayesian models. Indeed, sensory processes take place over a cascade of processing stages among hierarchical cortical areas. These areas are not only hierarchically organised, but they also present certain important asymmetries in their connections (see Friston, 2005, 2010). In particular, there seem to be *forward* connections running from lower to higher regions, and *backward* connections going the other way around. A possible functional interpretation of these asymmetries can be found in Friston's model of cortical hierarchies according to which backward connections transport information about the expected causes of the activities at the lower levels (i.e. priors), and forward connections play a modulator role by transmitting prediction-error information higher up in the hierarchy. Perception would then result from the interrelation of these forward and backward signals.

An additional source of evidence comes from studies on mental disorders, such as schizophrenia and psychosis. The key idea in the case of schizophrenia is that understanding its positive symptoms requires understanding the disturbances in the generation and in the precision of the prediction-error signals. The hypothesis (Corlett *et al.*, 2009; Fletcher & Frith, 2009) is that malfunctions in the working of the hierarchical models can yield continuous and persistent false prediction-errors, which then propagate all the way up into the hierarchy and, in severe cases, deeply modify and affect our model of the world. As a consequence, what should be experienced as improbable becomes the less surprising. Given that perception is influenced by a continuous cascade of top-down signals matching bottom-up signals, in these cases the cascade of misinformation reaches the lower-level processes, thus yielding false perception and wrong beliefs about the state of the world. In the case of psychosis, researchers have found that patients report a change in the intensity with which they perceive the world (e.g. acknowledging a much louder background noise or brighter colours — Corlett *et al.*, 2007, 2011) before becoming psychotic. Even normal everyday life experiences appear to be more vivid, novel and important (Kapur, 2003).

Among the symptoms of psychosis we can find: misrepresentation of reality, delusions, hallucinations and experience of one's action as under the control of external agents. Some have suggested that all these symptoms could find an explanation within the Bayesian framework.

Consider the case of delusion of alien control (Hohwy, 2004, 2013; Hohwy & Rosenberg, 2005; Fletcher & Frith, 2009), that is, the false belief that someone else is controlling our actions. This delusion can be seen as a result of anomalous prediction-errors. When I am engaged in self-generating actions, the precision-weighting on the relevant proprioceptive prediction-error must be set high. If the proprioceptive prediction-error is set high and my top-down predictions can resolve it, I feel that I am the agent of my own actions (Blakemore *et al.*, 2002). Problems arise when there is no match between predictions and actual proprioceptive inputs. In these cases, an agent's experience becomes surprising. These mismatches could depend on mistakes in the generation of prediction-errors or in their weighting. While the subject knows that she wanted to move and that she acted on that intention, the signals she receives, which are not attenuated, indicate that someone else made her move. A prediction-error that cannot get explained away by top-down signals emerges and gets propagated upwards in the cortical system. The system must now find another hypothesis that can account for the data, thus explaining away the prediction-error (e.g. "someone else made me move").

These malfunctions, which at first arise as rational responses to unusual situations (Hohwy & Rosenberg, 2005), can result in persistent and highly false prediction-errors that force, in severe cases, extremely deep revisions in our model of the world. In these severe cases, what should appear as improbable (e.g. persecution) becomes the less surprising cause of our sensory signals.

Within the Hierarchical Bayesian framework, hallucinations and delusions (and perception and belief), which are normally considered the result of two different processes, both involve a similar mechanism that allows top-down predictions to match sensory signals.

A Bayesian hierarchical model also makes explicit and testable predictions about the role of different neurotransmitters in signalling prediction-errors and their precision or uncertainty. There are studies indicating that dopaminergic activity might encode the degree of precision or uncertainty of certain prediction-errors, while errors could be carried by glutamatergic neurotransmitters (i.e. Corlett *et al.*, 2011). Much work still needs to be done, but, interestingly, the Bayesian neurocomputational framework seems to operationalise some of its central claims for future experiments.

6.6 – Neural representations

I have shown so far how (Bayesian) neurocomputational models can be used to make sense of some aspects of perception, decision-making and mental illness. In this section I will focus my attention on the nature of the internal neural representations, which are the building blocks of Bayesian inference.

Bayesian internal representations are understood as neural states that carry information about some variables in the word and about the past experience of the agent.

Neuroscientists usually identify representations in a certain brain region by working out how the response profiles of certain neurons (i.e. their patterns of action potentials, or spikes) connect to the agent's behavioural outcome and to the state of affairs in the world. Since neural activity is often noisy and variable, the informative aspect of neural responses can only be captured in terms of probabilities over a population of neurons. Single neurons don't represent features or micro-features of any environmental state; rather, populations of neurons might represent probabilities of the possible values of the stimulus.[29] The aim of the cognitive system is to infer the nature of the signal source on the basis of the probability distribution of the neural response and of the prior probability of the causal structure of the environment.

Neural representations could be individuated in terms of encoding-decoding mappings (Eliasmith, 2003). Neural encoding identifies the functional dependence of some neural property on some property of the stimulus. As argued above, action potentials could specify the neural encoding. Neural decoding, instead, refers to the process of inferring, or estimating, the property of the stimulus from the specific type of neural encoding, that is, from the property of some neural response. Physical features of neural populations' firing responses might count as the basic units of neural decoding. The estimated value of the stimulus that results from the neural decoding process is then used by the system to carry out the cognitive task.

The way in which the decoding process could estimate the value of the stimulus that led to a certain firing pattern depends both on the prior information of the system and on the likelihood of stimulus, that is, the estimation depends on the generative model that the system could use, and on how prediction-error is weighted. Given a certain neural pattern of activation, then, the decoding process estimates how likely it is that a certain stimulus is indeed in the environment by relying on the activity of predictions carried by the activity of neurons from higher to lower levels in the hierarchy.

Consider Hubel and Wiesel's (1962) experiment. They moved a bar of light at different angles across the region of the visual field where cells responded to light (i.e. the cells' receptive field) to find out whether there were cells representing features of the stimulus. They observed that the number of cell's action potentials

[29] As Eliasmith clearly puts it: "Neurons don't 'detect' things [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).

that fired depended on the angle of orientation of the bar. In other words, they discovered that the response tuning curves of the cells (i.e. a plot of the average firing rate of the neuron as a function of the relevant stimulus' feature) could be indicative of the orientation of the stimulus: the maximum average response of the cells corresponded to a specific angle orientation. They then called the angle that evoked the maximum average response "the preferred orientation" angle of the neuron.

Within the Bayesian neurocomputational framework, the content of an internal representation is then characterised by the neural turning curve, by the maximum average response and also by the way in which the representation is used within the system. The content can then be adjusted within the generative model via the interaction of top-down and bottom-up signals, which makes it highly dynamic and context-sensitive.

If we grant that spikes of neurons can represent basic physical features, such as the orientation of a bar of light, we can expect larger populations of neurons to encode more complex and abstract representations at higher levels in the cortical hierarchy (see Eliasmith, 2003). Lower-level representations would then be influenced and shaped by higher-levels ones while remaining highly sensitive to raw incoming sensory information, and higher-level more complex representations would depend on the lower-levels ones via prediction-error signals. This way, the sensory system could incorporate statistical dependencies between representations at different levels of complexity and abstraction.

Saying that a system carries out cognitive functions in an optimal or approximately-optimal way means that the system can take into account the uncertainty in the available information to maximise the probability of understanding what is in the environment. Understanding the cause of its sensory inputs is, then, required to appropriately achieve the desired result. Optimality is, therefore, not a fixed universal property; rather, it depends on the prior and on the measurement of the likelihood.

Referring to representations and representational content does explanatory work, as I will argue in the next chapter, because it shows "how the system connects with its environment: with the real-world objects and properties with which it is interacting, and with the problem space in which it is embedded" (Shea, 2013, p. 499).

Neural evidence shows that actual neural variation is less disparate than it might appear at first. fMRI studies, for instance, seem to suggest that there are similarities in patterns of activation across individuals and trials such that a same representation can be realised in similar ways across subjects and trials. This, in turn, opens up the possibility to infer an agent's psychological state from observations of certain brain properties and to predict, with reasonable accuracy, a subject's performance in similar tasks.

Discovering correlations between components of Bayesian algorithms and neural signals is important because:

- It points out that some kind of neural algorithm that computes over a specific component (or a similar one) is indeed realised in the brain (Mars *et al.*, 2012, p. 259)
- It indicates that, when a certain component is neurally represented, it is possible to discover some similarities in the specific activation pattern produced

As shown in the case of reward decision-making processes, fMRI techniques help to uncover the neural presence of prediction-error (or of a similar quantity) in neural circuits: specific BOLD signals seem to relate quantitatively to representations of prediction-errors in a quite linear way.

Although many studies are now trying to specify the nature, role and format of neural representations, the exact way in which they are learnt, encoded and updated through neural activity is, at the moment, largely unknown. Researchers have only recently begun to study the possible neural basis of Bayesian computations for relatively simple perceptual tasks (see section 7.5) and the way(s) the brain represents uncertainty and computes over it to perform Bayesian inference is still poorly understood. With respect to this, the Bayesian neurocomputational framework seems to offer some additional means to better understand the nature of internal representations. Leading figures of this approach (i.e. Griffiths *et al.*, 2010; Tenenbaum *et al.*, 2011) claim that Bayesian models allow researchers to explore the nature of representations, thus opening up the possibility for representational pluralism:

> "Probabilistic models […] provide a transparent account of the assumptions that allow a problem to be solved and make it easy to explore the consequences of different assumptions. Hypotheses can take any form, from weights in a neural network, to structured symbolic representations, as long as they specify a probability distribution over observable data. […] The approach makes no *a priori* commitment to any class of representations or inductive biases, but provides a framework for evaluating different proposals." (Griffiths *et al.*, 2010, p. 358)

There is much ongoing work devoted to uncover which algorithms the mind does use and how they are realised in neural circuits. Some studies indicate that the brain might use Monte Carlo or stochastic sampling-based approximations to approximate optimal Bayesian statistical inference.

Another big obstacle consists in understanding if and how structured symbolic knowledge, which is often considered essential for certain forms of cognition and thought, can be represented in the brain. In contrast to connectionism, which sidesteps these challenges by denying that the brain can actually encode this kind of knowledge, the Bayesian neurocomputational framework leaves room for the possibility that the brain might compute over more structured symbolic representations. These representations, however, would not be rigid, static or hard-wired; rather, they would grow dynamically in response to noisy data from the environment while remaining embedded in hierarchical generative models (e.g. Tenenbaum *et al.*, 2011).

6.7 – Conclusion

In this chapter I examined the structure and the methodology of the neurocomputational framework of explanation mostly by analysing a family of models — the Bayesian models — that belongs to it.

By discussing various examples of these models in practice, I showed that neurocomputational explanations are a special kind of subpersonal explanations that aim at explaining cognitive behaviour in terms of the various ways in which the brain, which is seen as a processor of information that traffics in representations, copes with its environment. In particular, neurocomputational explanations look for correspondences between stages of information processing and biological transactions among neural populations.

I focused on various cognitive behaviours that appear to result from some sort of prediction-error minimisation process and I showed why this process seems to be a central building block of a mechanism that allows agents to perceive what is in the environment, to learn how to predict the consequences of their behaviours and to perform in a nearly-optimal way. I then discussed various empirical data that speak in favour of the existence in the brain of some quantities similar to the prediction-error and to the stimulus' expected values that appear to play important roles in various cognitive behaviours.

Chapter 7 - The Explanatory Virtues of the Bayesian Neurocomputational Framework

7.1 – Introduction

A good explanation of cognitive behaviour needs to be predictive and mechanistic. This is the claim that I made at the beginning of the thesis and that has accompanied our journey so far.

In this chapter I will examine whether the neurocomputational framework can provide explanations of cognitive behaviour that are better than those offered by the folk-psychological, the anti-representational and the physiological subpersonal frameworks.

I will argue that the neurocomputational framework can better account for cognitive behaviour and I will highlight the features that make it different and superior with respect to the other three frameworks.

I will claim that, in contrast to the folk-psychological and the anti-representational frameworks, the neurocomputational framework successfully meets the predictive criterion and offers useful means to arrive at a full mechanistic identification of the responsible process(es) underlying cognitive phenomena. In particular, I argue that predictions play a central role in the neurocomputational framework and that the framework's openness to an analysis of the possible implementation of cognitive processes together with the growing operationalisations of some of its central claims make it the most adequate framework to explain various aspects of our cognitive life. In addition to this, I will argue that, in contrast to the purely subpersonal framework, the neurocomputational framework offers explanations of cognitive phenomena that incorporate both personal and subpersonal information about their responsible processes.

7.2 – Predictions

Predictive power is central in the neurocomputational framework and it is used to evaluate the goodness of explanations.

The important role attributed to prediction is first of all reflected in modelling design: neurocomputational models are designed to avoid the risk of being too sensitive to the peculiarities and noise characteristic of a given data set, which are unlikely to get generalised to other cases. One way in which models avoid being too sensitive to a given set of data is by being simpler (e.g. by containing fewer parameters) than other models used to account for the same data (see e.g. Chalk *et al.*, 2010; Weiss *et al.*, 2002).

Despite the centrality of predictions, the choice of ad-hoc parameters (e.g. priors and likelihoods in the case of Bayesian models) can limit neurocomputational models' predictive power. If parameters are selected for their mathematical tractability (e.g. flat prior, Gaussian distributions, and so on) rather than for their empirical adequacy, models don't incorporate real aspects of the biological mechanisms responsible for the phenomena under study. Assuming that people operate on the basis of flat priors, for instance, consists in assuming that people approach cognitive tasks without any prior expectation. This is an implausible assumption: people usually approach contexts and tasks that are, in some sense, similar to others they have previously encountered, and, even when they approach new scenarios or face new tasks, expectations from their evolutionary and developmental history might be in place, thus affecting their performances (e.g. perceptual and motor systems are often already geared towards certain responses in specific contexts).

If the predictive success of neurocomputational models depended solely on mathematical tractability, we wouldn't have reasons to conclude that these models shed light on the real mechanisms responsible for cognitive performances. If this were the case, then Bayesian explanations would not be so different from folk-psychological or anti-representational explanations. As I have shown in previous chapters, folk-psychological explanations are often more useful to predict behaviours than to explain them, and dynamical and anti-representational explanations can predict how a system evolves through time, but not why it has a certain capacity in the first place.

The claim I want to make in this chapter is that the neurocomputational framework can offer more than just predictions: by making "good" predictions, the Bayesian framework aims at providing mechanisms.

7.3 – The search for mechanisms

Researchers working within the neurocomputational framework are becoming increasingly aware of potential limits of the framework's modelling design: models can explain cognitive behaviour only to the extent that they incorporate ecological and biological considerations in their construction. When models are adequately constrained, they can yield secure and informative predictions, and, ultimately, explain cognitive phenomena.

To this end, some investigators have started designing psychophysical experiments that can indicate the information subjects really have when performing in certain tasks (e.g. Maloney & Mamassian, 2009). Knowing this information means being able to predict how subjects will perform in similar tasks. Let me briefly discuss this kind of experiments.

As I have examined in the previous chapter, according to the Bayesian neurocomputational framework, subjects face a task with a certain hypothesis space and priors. In the case of Hierarchical Bayesian models, the origin of the hypothesis space and priors is addressed by positing not just one single level of hypotheses, but multiple levels, each one generating a probability distribution on variables at the level below. The hypotheses and priors required for a specific task are learnt by the system via Bayesian inference across levels. Once the hypothesis space and the priors are learnt, a subject can perform a cognitive task by computing the correct posterior distribution via Bayesian operations. The acquired posterior distribution will then constrain, in the form of a new prior, the subject's future performance in the same or in similar tasks. The subject is therefore expected to perform better and faster in tasks where this new prior is required.

Consider the case of perception. Imagine a subject who is learning how to perform two different tasks. In Bayesian terms, we would say that the subject is learning two combinations of priors and likelihoods that can account for the states of affairs in the world. Now, if perception is Bayesian, the subject will be able to use the knowledge acquired in the two perceptual tasks to perform in a new task that requires the combination of the previously encountered priors, likelihoods and posteriors. If the subject's performance in the new task is close to optimal and in accordance with the model's predictions without much practice, then we will have evidence that the subject performs on the basis of internal representations in a Bayesian fashion.

This form of "transfer learning", which has already been useful in machine learning and artificial intelligence, is critical for humans as well (Tenenbaum *et al.*, 2011). Transfer learning is a methodology that allows us to envisage a future where (Bayesian) neurocomputational models will make predictions about behavioural performances in psychophysical tasks based also on knowledge of the internal process (and its components) that leads a subject to behave in a way rather than in another. These predictions will be more secure than those based solely on psychophysical data.

Whether neurocomputational models can provide this kind of prediction for humans is still an open question. It might turn out that the quantities needed for a subject to perform in a certain task cannot be easily generalised to novel tasks. This, however, wouldn't affect the explanatory goodness of the overall neurocomputational framework. The capacity of the framework to suggest predictions that can be experimentally tested to help uncovering the real components and their interactions responsible for certain cognitive performances is enough for us to consider the neurocomputational framework as more fruitful than the others in advancing our understanding of cognition. Indeed, its model-based approach provides tools to discover the unobservable nature of internal mechanisms for cognitive behaviour, by suggesting predictions, by operationalising claims, and by performing experiments to confirm or disconfirm them.

The cue integration study (Ma *et al*., 2006), which was originally motivated by psychophysical results, can be of help to clarify the point. This study aimed at uncovering the neural basis of subjects' Bayesian nearly-optimal performance in cue integration tasks. By relying on neural data, Ma and colleagues suggested that the Poisson-like variability of certain cortical neurons allow a network of neurons to carry out cue integration using linear operations on population activities. This interpretation, far from being merely a description of the mechanism underlying cue integration, was important to generate novel and potentially informative predictions concerning features of neural activities and specific organisations of neural circuits that could turn out to be necessary to perform cue integration in a Bayesian nearly-optimal way. Ma and colleagues could, for instance, predict that if subjects performed in a Bayesian fashion in cue integration tasks and if neurons have a Poisson-like variability, then we should expect the sum of the activations of these neurons to be equal to the response of multisensory neurons. If correct, these predictions would be informative (i.e. they would uncover the nature of some aspects of the mechanism of which we were unaware) and secure (i.e. they would be based on reliable, well-evidential grounds).[30] In particular, they would be secure as long as they would depend on both psychophysical studies (i.e. identification of the particular types of circumstances where people behave as ideal Bayesian observers) and on some features of neural circuits (e.g. Poisson-like variability).

7.4 – Pay-offs of the Bayesian neurocomputational framework

So far, I have argued that:

- Predictive power is central in the Bayesian neurocomputational framework
- Predictability is used to evaluate the goodness of explanations
- The Bayesian neurocomputational methodology can be used to identify the mechanisms responsible for cognitive performances

What about the Bayesian neurocomputational framework's explanatory goodness, then?

Throughout the thesis I asked the same questions with respect to the folk-psychological, the anti-representational and the purely subpersonal frameworks. Let me briefly recapitulate the results of my analysis so far.

In chapter 1, I tried to understand why folk-psychological explanations, which are based on beliefs and desires (and their connections), are often predictive, and I concluded my analysis claiming that the folk-psychological framework cannot

[30] Secure predictions are based on solid, reliable grounds. A model can yield secure predictions when it specifies under what circumstances a phenomenon is likely to obtain. If a model identifies under what conditions, in virtue of which components and relationships among components a phenomenon is to be expected, then the model offers information about a candidate mechanism.

provide good explanations of cognitive behaviour. By analysing belief-desires models of explanations, I showed that the central notion of cause that they employ is problematic and I highlighted how this negatively affects our ability to distinguish rational re-descriptions from real explanations within the folk-psychological framework. In particular, I claimed that in order to justify that certain beliefs and desires cause a behaviour it is not sufficient to show that hadn't they occurred, the behaviour wouldn't have occurred either. Counterfactuals statements can be used as evidences for the existence of certain causal relations, but they can't establish the truth of causal claims. Related to this, I argued that another problematic aspect of the folk-psychological framework is the purely functional characterisation of the causes of cognitive behaviour, and I suggested that causes need to be characterised both functionally and structurally. In chapter 4 I analysed folk-psychological normative explanations and I showed that they need to be supplemented by information from lower levels of analysis, both to uncover what is constitutive of personal phenomena, and to explain, rather than redescribe, cognitive behaviour.

In chapter 2, I examined the anti-representational framework. According to this framework, it is possible to explain cognitive behaviour by studying the way in which cognitive systems interact with their environment. Anti-representationalists claim that a good explanation of a cognitive capacity is possible when brain, body and world are considered as a unique system that changes through time. I showed how this assumption makes the framework unable to account for why a system has a certain capacity in the first place. I highlighted some problems related to: (i) the use of lumped parameters that can't be mapped onto any biological component; (ii) the weak relationship of instantiation between mathematical models and systems that doesn't allow the identification of the responsible processes underlying cognitive behaviour.

In chapter 4 I analysed the purely subpersonal framework and I argued that it cannot provide good explanations of cognitive behaviour because folk psychology doesn't simply play a heuristic role in driving research, but it often plays a constitutive role in explaining cognitive behaviour. An adequate explanation of cognitive behaviour requires both the personal and the subpersonal level of analysis.

In what follows, I will provide some reasons for why we should be optimistic about the explanatory pay-offs of the neurocomputational approach. If I can show that the Bayesian neurocomputational framework can offer an account of how subjects performs cognitively that doesn't have, or that can overcome, some of the other frameworks' limits, then I will be justified in concluding that this framework is the most apt to generate good explanations of cognitive behaviour. In particular, I will try to answer the following questions:

- Can the Bayesian neurocomputational framework suggest a better account of cause?
- Can the framework provide a better analysis of the relationships between different levels of analysis?

7.4.1 – Functions and structures

The methodology adopted within the Bayesian neurocomputational framework is based on the assumption that, if a system has a certain property, that property depends on the nature and on the organisation of its component parts. This methodology is different from that of folk psychology where explanations are based solely on behavioural data and little attention is devoted to the study of the inner workings of the brain. The neurocomputational methodology is also different from that of anti-representationalism that identifies lumped parameters independently from their biological counterparts.

Working within a neurocomputational framework implies a deep study of the functional (e.g. the ability to carry a certain type of information transaction) and structural (e.g. neural type, anatomical position, and so on) features that characterise components and processes responsible for cognitive phenomena. Structural and functional characterisations are here seen as partial and complementary in the sense that each one draws on specific properties of the other: to get a comprehensive mechanism, activities must be localised in parts so that working parts can be established. This aspect of the neurocomputational methodology gets particularly clear if we consider the studies designed to discover the mechanism underlying reward-guided decision-making.[31]

The case of reward-guided decision-making is a good example of how a model-based strategy, as the one offered by the neurocomputational framework, can help to specify unobservable components and important features of neural mechanisms.

Neuroscientists studying habitual decision processes have shown that there are certain neural circuits that work in ways that resemble those of the temporal-different (TD) learning mathematical model. In particular, they argue that it is possible to establish a correlation between the reward prediction-error in a TD learning model and a BOLD signal in specific neural circuits. This signal, which corresponds to a specific activation of dopamine neurons, is understood as playing a crucial role in learning. Accordingly:

> "[…] when a correlation is found between a model components and a neural signal, that is taken as evidence that the brain implements an algorithm that involves calculating over that component." (Mars *et al.*, 2012, p. 256)

These data, despite not being the ultimate evidence that the brain is computing over internal representations and prediction-errors by employing a TD learning algorithm, at least suggest that some quantities of that algorithm are realised in the brain and are computed over to enable the system to perform appropriately.

The methodology adopted by the neurocomputational framework is full of potential for uncovering the mechanisms underlying cognitive phenomena, thus yielding good explanations. Despite being only recently adopted to study the brain, it

[31] For more details on reward-guided decision-making, see chapter 6.

has already been proven useful. This gives us sufficient reasons to believe that, in the future, its strategy will “yield conclusions about the class of algorithms that it is likely that the brain uses in performing a given task, including identifying neural structures that are involved in representing some of the quantities over which the algorithms compute” (*ibid.*, p. 258).

7.4.2 – Top-down approach

A second reason why we are justified in considering the Bayesian neurocomputational framework superior to the other frameworks is that it employs a top-down approach to the analysis of how cognitive systems perform in various tasks. Thanks to this approach, the framework can allow testable predictions, leave room for exploring a broad range of different assumptions about how cognitive systems might perform certain cognitive behaviours and open up the possibility for representational diversity.

The neurocomputational top-down approach starts with defining possible ways in which systems can perform cognitively, and then generates experimentally testable predictions (e.g. Griffiths *et al.*, 2010; Pellicano & Burr, 2012) to confirm or disconfirm the hypotheses.

As a matter of fact, there are “myriads of ways in which human observers behave as Bayesian observers” (Knill & Pouget, 2004). This has fundamental implications for neuroscience, particularly for how we conceive of neural computations and the nature of neural representations of perceptual and motor variables. Within the Bayesian neurocomputational framework, for instance, the fact that background knowledge is encoded in probabilistic generative models doesn’t mean that the hypotheses constituting this background knowledge need to be in a single format. Rather, by operating on a broad range of possible formats of representations, Bayesian models can search and evaluate different proposals within the same type of explanatory framework. A model can be defined and representations in a particular format specified. If the model doesn't fit the behavioural data, a different model with representations in a different format can be suggested and tested (Kemp & Tenenbaum, 2008).

Leaving room for assessing different hypotheses and for representational diversity are two positive features of the neurocomputational approach that show how the search for mechanisms can be potentially free from too strong *a priori* commitments. In this sense, hypotheses and priors can take any form, from weights in a neural network to structured symbolic representations, which are those that might be involved in the most complicated and demanded cognitive tasks.

7.4.3 – Answers to the interface problem

A third reason to prefer the neurocomputational framework is that it offers interesting means to understand the relationship between personal and subpersonal-levels of analysis under a new light. This is a clear advantage with respect to the folk-psychological, the anti-representational and the purely subpersonal frameworks, and it is particularly crucial now that scientific disciplines studying the workings of the brain are rapidly growing.

In chapter 4 I discussed the relationship between personal and subpersonal explanations by adopting the lens of the interface problem, that is, the problem of how we should relate explanations couched in different vocabularies and belonging to different levels of analysis. I followed Bermudez's (2005) in defining the interface problem as a problem about the relationships among various disciplines of study, such as folk psychology, scientific psychology, cognitive science and neuroscience. I argued that solely personal-level explanations and solely subpersonal-level explanations are not suitable to explain cognitive behaviour, and I concluded that a better understanding of the relationships between levels of explanation is needed. In this section I will explore the position of the neurocomputational explanations with respect to the interface problem.

The peculiarities of this approach that, I believe, can clarify its position with respect to this problem are the followings:

- It employs a top-down, function-first methodology
- It adopts both the vocabulary of the brain (e.g. neurons, activations of neurons and neurotransmitters) and notions that commonly belong to the personal-level vocabulary (e.g. belief, expectation, internal representations, rationality and inference)
- Some studies within this framework attempt to uncover something about the nature of the processes through which agents come to behave in a way that is Bayesian-rational

I believe that these features make the neurocomputational framework a good starting point to shed new and potentially interesting light on the interface problem. Let me clarify this claim.

The framework incorporates quite naturally the methodological stance according to which neuroscience should ultimately offer some contribution to the way in which we think about the basic phenomena of the mind (see chapter 4). Employing a top-down approach means that, in order to allow neuroscience to say something about cognitive phenomena, we should first of all observe the phenomena and then hypothesise ways in which cognitive systems might perform them. This first step (i.e. a computational analysis for a specific cognitive task) is necessary to discover something potentially informative inside the brain. Having a top-down methodology allows the exploration of the possible connections between the functional level and the level of the brain because, with a description of the task at hand, we can consider which processes could approximate the required computations and then investigate the kinds of neural components and neural interactions needed for those approximations to be carried out.

Accordingly, the neurocomputational framework seems suitable for the co-evolution among levels that Patricia Churchland suggested (1986), that is, the idea that information at different level of analysis can interact and constrain each other. I have already discussed why co-evolution is a good strategy to understand cognitive phenomena previously. Here I intend to stress how the neurocomputationalism can allow it.

Someone might wonder whether the neurocomputational framework adopts a reductionist or materialist methodology: if scientists can tell us how cognitive processes are realised in the brain, then cognitive explanations won't require personal-level folk-psychological notions anymore. I want to argue that this is not, and it doesn't have to be, the methodology adopted within the Bayesian neurocomputational framework for even once we do have an implementational description of a cognitive behaviour, notions such as that of expectation and internal representation will still be required to make the cognitive behaviour intelligible. Indeed, it is peculiar to this approach the idea that cognitive behaviours are possible because cognitive systems can deploy internal generative models as surrogates of some aspects of the environment. These internal models are made out of internal representations, whose transformations allow cognitive systems to behave appropriately. The notions of internal models and internal representations are, therefore, central in neurocomputational explanations. What the framework can offer is, rather, a better (structural and functional) specification of the nature of these notions. As Patricia Churchland (2004, p. 49) claims "these discoveries begin to forge the explanatory bridge between the experience-dependent changes in neurons and the experience-dependence guidance of behaviours".

From the personal-level of explanation, the (Bayesian) neurocomputational framework inherits, for instance, the notions of expectation and inference. Bayesian studies also show the existence of a much closer correspondence between optimal statistical inference and everyday cognition than commonly supposed: the brain seems to approximate quite neatly in fundamental aspects of its operations a certain kind of ideal, that is, the Bayesian rational ideal. This suggests that the Bayesian top-down approach can shed light on how practical rationality might be enabled:

practical rationality is possible thanks to predictions made on internal representations and generative models in accordance with a rational norm — that of maximisation of expected utility. Working within the neurocomputational framework, then, allows us to say something about why — and not only how — people behave in certain tasks in a Bayesian-like fashion.

To sum up, the neurocomputational framework allows the personal level and the subpersonal level to interact in multiple ways:

- The personal level guides and motivates the search for mechanisms and components in the brain and it offers important conceptual tools to understand the nature of cognitive phenomena
- The subpersonal level offers grounds to justify personal-level claims, to uncover something about the constitutive nature of certain personal-level phenomena (e.g. mental illnesses, rationality and inference) and to identify the neural mechanisms responsible for them

Consider the case of delusions analysed in chapter 6. Delusions are usually considered to be results of malfunctions in a putative belief-formation mechanism. Delusions are commonly distinguished from hallucinations, which are, instead, seen as consequences of breakdowns in the mechanism responsible for perception. Although there remain important differences between perceptual anomalies and delusions, the (Bayesian) neurocomputational model suggests an interesting link between perception and belief-formation mechanisms: they both involve the attempt to match the incoming sensory stimuli with top-down predictions about the causes of those stimuli.

Whether or not delusions and hallucinations result from the operation of very similar mechanisms, these studies show that working within the neurocomputational framework allows us to partially answer important constitutive questions about the nature of certain cognitive phenomena.

By suggesting a new, multilevel and model-based account of the interactions between inferences, expectations and learning, we have reasons to hope that this framework will, one day, offer a better understanding even of our own agent-level experience than that afforded by folk psychology.

To summarise, the neurocomputational framework has the following explanatory pay-offs:

- It values predictive power, which is adopted to evaluate the goodness of explanations
- It has the potential to characterise the causes of certain cognitive phenomena both functionally and structurally
- It leaves room for exploring a broad range of different assumptions about how people might solve certain tasks and it opens up the possibility for representational diversity
- It aims at identifying neural mechanisms, that is, the components and their regular interactions responsible for cognitive performances
- It can shed new light on the interface problem by letting the personal and the subpersonal levels interact in a fruitful way

7.5 – Neural representations and behavioural intelligibility

So far I have offered some reasons for why we should be optimistic about the explanatory pay-offs of the neurocomputational framework. In this section I will highlight that such explanatory purchase is also due to the central role that a special notion of representation, that of neural representation, plays in it.

My argument will be twofold. I will first show why neural representations are necessary for good predictions and I will then discuss why these representations can be considered real components of mechanisms responsible for cognitive performances. The neurocomputational framework, I will claim, allows a better specification of the notion of representation and of it role in cognition.

Within the neurocomputational framework, neural representations are understood as neural states that carry information about some variable in the world and that retain information about the past experience of an agent. Neural representations enter into causal processes in ways that depend on their physical properties.

At the beginning of the chapter, I argued that the ultimate goal of the neurocomputational framework consists in establishing bridges between the personal (and functional) level and the subpersonal (and neural) level of analysis. By connecting the two levels, the framework aims at making the functional and the neural vocabularies symmetrical. For this reason, neurocomputationalists try to understand how representations are encoded by neural activities and transformed by neural operations.

In section 7.2 I showed that the search for bridges between levels is motivated by the need to gain not only predictions, but *good* predictions of people's behaviour in cognitive tasks. By suggesting good predictions, I claimed, Bayesian models can aim at explaining cognitive phenomena in a genuine sense. I then suggested that predictions can be good (i.e. secure and informative) if they are based on knowledge of the brain too. By examining various applications of Bayesian models and empirical data, I claimed that there are reasons to believe that, in the future, neurocomputational models will make predictions about behavioural performances that will be secure because they will be based not only on behavioural data, but also on empirical data concerning the internal workings of the brain. In other words, good predictions will be possible when neural representations will be characterised both functionally and structurally: the level of the brain offers further evidence, control and testability that the system is indeed operating on a certain process and on certain components. This, in turn, allows more informative and precise confirmations (or disconfirmations) of possible explanations. The adoption of the methodology of "transfer learning" (see section 7.3) is a starting point in this direction. The identification of the information that subjects really have when performing cognitive tasks provides an important further source of evidence that can tell us whether subjects perform on the basis of internal representations in a Bayesian way. Transfer learning indicates, then, the beginning of a process that will uncover more and more about the internal nature of the mechanisms that underlie our cognitive abilities.

The fact that important components of putative mechanisms can be characterised both structurally and functionally is probably most evident in the work on reward-guided decision-making. This work provides one of the best cases of convergence between a functional description and a description of what goes on in the brain: investigators found co-variation between the amount of reward prediction-error predicted by the model and the BOLD signal in certain neural circuits of subjects performing on the same task (see chapter 6).

The rise of spatially-detailed imaging techniques (e.g. fMRI, PET) has indeed opened up the possibility to identify how certain functions are realised in the brain. There are already examples showing that there is a rather consistent effect on the measurements of BOLD signals, which are coupled to differences in neuronal firing rates, when a subject is representing the same content on different occasions (Mukamel *et al.*, 2005). These preliminary results indicate that a same representation might get implemented in the brain in a similar way across individuals and trials. If this were the case, we might, one day, be able to generalise and predict behavioural performance on the basis of subjects' neural activities.

Attributing content to certain neural activations is therefore required to explain the ability of an agent to generalise her correct performance to new tasks and to make sense of why a subject performs in a specific task in a way consistent with the performance of a Bayesian observer. I have previously shown that the existence of internal generative models, whose components are internal representations, is essential to account for how subjects can experience the world and not just sense data: cognitive agents perceive the world by meeting the incoming sensory signals with a top-down cascade of representing interacting causes. The use of internal knowledge has several advantages: it enables us to hear what is said despite noisy surroundings, to adjudicate between alternative possibilities each one consistent with the stimuli, and so on.

If this is the case, then, we are justified in saying that neural representations are real in a very specific sense: they are real because they are explanatory inevitable when the explanandum phenomenon is formulated in functional or representational terms. In other words, neural representations are real elements of subjects' internal models because invoking them is necessary to explain why they can perform in the ways predicted by the neurocomputational models. Invoking neural representations shows how cognitive systems are connected with their environments and with the tasks in which they are embedded (Shea, 2013).

This claim is in line with the conclusion that I have drawn in chapter 3 with respect to William Ramsey's partial eliminativist proposal. There, I argued that we should understand a system as representational when there are enough reasons to do so, even in the absence of a full-blown theory of representation. If we can show that we can predict and generalise a cognitive system's behaviour by attributing representations to it, then the system is employing representations: representations are genuinely real components of the mechanism that the system uses to perform successfully in its environment. Identifying mechanisms is, therefore, a necessary step to generate good cognitive explanations, which are explanations that need to account both for why cognitive systems behave in certain ways and for how they do so.

Accordingly, even if we constrain our theory of representation on the basis of explanatory usefulness, this is still consistent with representations being real internal entities. They are real, even if it is not clear yet how exactly they are realised within the system.

When representations are interpreted in this way, they are a departure from more classical views on representations:

- They don’t depend solely on the sensory inputs from the environment; rather, they are also influenced by inputs from within the brain, and, in particular, by inputs from other cortical areas
- They are not static. The neuronal responses following a certain stimulus may vary according to the context and to the background knowledge a subject can bear on the task (in Bayesian models, this knowledge is in the form of the current winning top-down prior prediction). In this sense, even if the system makes use of structured symbolic representations, these representations don’t need to be rigid or hard-wired, but can grow dynamically in response to noisy data from the world

Whether or not we will find one-to-one mappings between functional states and neural states, the study of the brain — motivated by the goal of finding bridges between the personal and the subpersonal levels — will still be useful to gain a better understanding of our mental life. As claimed above, by suggesting a model-based and multilevel account of the interactions between inferences, expectations and learning, the framework could, one day, offer a better understanding even of our own agent-level experience than the one afforded by folk psychology.

If mappings between functional states and states in the brain are identified, the framework will be able to offer predictive and mechanistic explanations of cognitive behaviour. The personal level of analysis will still be necessary to account for cognitive phenomena when these will be formulated in functional and representational terms. The subpersonal level of analysis, instead, will be adopted to explain phenomena formulated in neural or physiological terms or to achieve different epistemic or practical goals. The computer analogy might be of help here. The functional level of analysis is necessary when our aim is to program the computer, while the implementational level is required when we need to intervene on it to fix some problems.

Working within the neurocomputational framework, then, allows the construction of models that can be better confirmed by empirical results with respect to the folk-psychological and anti-representational ones. If the best models, and, consequently, the explanations they can achieve, are those that can be best confirmed and justified through evidence, then the neurocomputational framework is the most apt to genuinely explain cognitive phenomena by integrating personal and subpersonal information about their underlying processes.

7.6 – Conclusion

In this chapter I argued that we can make progress in understanding cognitive behaviour by adopting the neurocomputational framework. I claimed that, although we don't have any fully worked out mechanistic explanation at present, we have sufficient reasons to believe that neurocomputational explanations will be good explanations, that is, predictive and mechanistic.

I argued that a central component of neurocomputational explanations is the notion of neural representation. Such notion, that still needs to be properly understood, is necessary to account for how a cognitive system can approach the world, handle its uncertainty and perform cognitively. In particular, I showed that neural representations are necessary for an explanation of cognitive phenomena to be predictive and mechanistic. When representations are defined both functionally and structurally, the resulting neurocomputational model can generate more secure predictions. Neural representations are also necessary for an explanation to be mechanistic because they are the components over which inferences can be performed. In this sense, I argued that the notion of cause employed within the neurocomputational framework is more precise than that of other frameworks previously analysed. Accordingly, I showed how the Bayesian neurocomputational framework values the search of neural mechanisms by incorporating ecological and biological considerations in the modelling design. This allows researchers to test predictions not only about behavioural performances, but also about internal mechanistic features.

In addition, I stressed the importance of the model-based top-down strategy as a means to explore a broad range of assumptions and hypotheses about the possible mechanisms underlying cognitive abilities. This advantageous methodology is also capable of shedding new light on the explanatory interface problem by uncovering something about the constitutive nature of certain mental phenomena.

For all these reasons, I conclude that working within the neurocomputational framework can help us to make progress in our understanding of cognitive behaviour.

I would like to stress that the goodness of the neurocomputational framework doesn't depend on the success of a model or of a family of models that belongs to it. My analysis of Bayesian models has been instrumental to a broader discussion over the structural, methodological and explanatory features of the neurocomputational framework. In this sense, whether Bayesian models will explain the mechanisms underlying cognitive abilities or not, whether the brain does implement Bayesian inference and whether internal representations do encode probability distribution are issues that are not directly relevant to the main conclusion of the thesis. The project has aimed at identifying certain features of explanatory frameworks that could generate adequate explanations in cognitive science. With regard to this goal, I have argued that good explanations of cognitive behaviour need to be predictive and mechanistic. I then indicated a framework — the neurocomputational framework — that can allow the search for this kind of explanations better than others. My conclusions are, therefore, not directly affected by the empirical success of Bayesian models. Rather, they depend on a different sense of "success" of the neurocomputational framework. For it to be able to identify proper mechanisms underlying cognitive behaviour, it needs to provide scientists with, for instance, a much clearer notion of what it means for an algorithm or for a specific quantity to be realised in the brain. The lack of a clear understanding of neural realisation negatively affects the goodness of the overall framework and, in turn, the experiments and the interpretations that scientists are allowed to carry out. On the positive side, the framework clearly strives for such clarification, as I have exemplified in my analysis of various cognitive behaviours.

In conclusion, despite the open questions and the specifications that are still required, adopting the neurocomputational can make progress in our understanding of cognitive behaviours and their underlying processes.

Conclusion

Various disciplines of study are devoted to understanding the processes underlying cognition, but there is still little consensus on the features that distinguish adequate from inadequate explanations of cognitive behaviour. There are currently four major frameworks that try to explain cognitive behaviour: the folk-psychological, the anti-representational, the subpersonal physiological and the neurocomputational frameworks. The goals and standards adopted by investigators working within these different frameworks are, however, largely lacking explicit articulation. This makes it difficult to understand whether these frameworks offer incompatible rather than complementary attempts to explain aspects of our cognitive life, thus limiting the progress in this field.

I opened this thesis with the following questions and answers:

Q_1: Which norms and values are used to construct, evaluate and justify models and explanations in cognitive science?

> A_1: Currently there are at least four different frameworks that try to explain cognitive phenomena. Each of these frameworks adopts different values and standards.

Q_2: What are the necessary desiderata of good explanations of cognitive behaviour?

A_2: A good explanation in cognitive science should be predictive and mechanistic.

Q_3: How can we make progress in our understanding of cognitive behaviours?

A_3: Adopting the neurocomputational framework is one way to make progress in our understanding of cognitive behaviours and their underlying processes.

I wish to conclude by showing the questions and answers in light of the claims addressed and defended in the chapters.

The first claim defended in this thesis is that the four frameworks pursue different explanatory goals and adopt different standards to evaluate the adequacy of cognitive explanations. Chapter 1 shows that the folk-psychological framework values predictability. Cognitive behaviour is explained by a generalisation of the form "if a person A desires B and believes that by doing C she will get B, then, *ceteris paribus*, she will do C". Chapter 2 discusses the anti-representational framework and argues that its explanatory goals are predictability and unification. Anti-representational models aim at predicting systems' behaviours — specifically how they vary through time — by importing theoretical, methodological and descriptive tools from other sciences. Chapter 4 examines purely folk-psychological rational explanations and purely physiological subpersonal explanations. The first part of the chapter shows how rational explanations aim at making behaviour intelligible. The second part of the chapter analyses solely subpersonal explanations. Explaining cognitive behaviour in purely neural physiological terms is shown to be the main goal of these explanations. Chapter 6 introduces and discusses the neurocomputational framework, which is based on the assumption that, if a system has a certain property, that property depends on the nature and on the organisation of its component parts. Chapter 7 argues that the neurocomputational framework aims at both predictability and identification of mechanism. Indeed, the framework values the ability to predict as a mark of a good explanation and it is open to the analysis of the implementation of cognitive processes made possible by the operationalisations of some of its central statements.

The second claim advanced in this thesis is that good explanations of cognitive behaviour need to be predictive and mechanistic. Predictability is a necessary feature of an adequate explanation because a good explanation has to provide us with information about the explanandum phenomenon that we could not have before. We need, for instance, to know that, given specific conditions, we should expect a certain phenomenon. However, predictability is not, by itself, a sufficient criterion to distinguish adequate from inadequate explanations; rather, it needs to be complemented by the identification of mechanisms.

Chapter 1 starts by discussing the major positions on the nature of scientific explanation and begins justifying why the ability to predict and to identify mechanisms are two necessary desiderata of adequate explanations of cognitive phenomena. It argues that a predictive description is not necessary also a good explanation by drawing on the well-known critique to the deductive-nomological model of explanation (e.g. Salmon, 1984). The identification of mechanisms, which requires both functional and structural analysis, is here suggested as a way to distinguish predictive descriptions from good explanations of cognitive behaviour. Chapter 3 provides further arguments for why the ability to predict is insufficient to validate the goodness of an explanation. The chapter also shows that the identification of mechanisms is necessary to better specify the relationship between models and modelled systems. Setting up correspondences on the basis of predictions between numerical sequences contained in the model and those of the real system's data is insufficient to explain these data. Rather, mathematical variables need to be identified in the physical substrate of the system for them to have real counterparts in the system performing a certain task. At the same time, revealing, as anti-representational dynamical models do, the existence of widespread patterns that apply to various physical systems does not bear on whether these models explain the phenomena or not. Chapter 4 argues that, without further evidence provided by the description of the responsible mechanisms, purely rational explanations at the personal level of analysis run the risk of being mere hermeneutic descriptions. The chapter also shows that solely physiological subpersonal explanations cannot adequately explain cognitive phenomena because they require the employment of certain personal-level notions (e.g. knowledge, information) to properly explain cognitive behaviour, and they need to start with a functional description of the cognitive process under study. Chapter 5 provides further arguments in favour of the identification of mechanisms by examining Jose Luis Bermudez's tripartite account of rationality (e.g. Bermudez, 2003). The chapter shows that the analysis of cognitive performances based on external behavioural criteria has to be supplemented by a deep study of how information is encoded and manipulated inside the brain. This information is necessary to confirm or disconfirm possible explanations.

Drawing upon this descriptive and normative analysis, it is argued that progress with respect to our understanding of cognitive behaviour is possible thanks to empirical discoveries, mathematical advances and also to the adoption of a framework that can play a genuine heuristic role. The third claim defended in this thesis is that cognitive behaviour can be effectively understood within the neurocomputational framework, which aims at identifying the workings of the mechanisms underlying cognitive performances. Mathematical and conceptual tools from statistical decision theory and reinforcement learning have been increasingly used to account for data concerning the neural basis of various cognitive behaviours, ranging from perception to action, to decision processes. The simultaneous reliance on theoretical and empirical approaches has allowed investigators to address more complex empirical questions about how cognitive systems can perform certain behaviours in more reliable and precise ways.

In chapter 7 these motivations were addressed in support of the claim that our understanding of cognitive processes can advance by working within a neurocomputational framework. This framework plays an important heuristic role in providing guidance to formulate novel empirical questions and to test current conceptual schemes. Looking for bridges between functional and neural explanations and aiming at making the functional and the neural vocabularies symmetrical by assessing the existence of correlations between mental states and neural states are, indeed, useful strategies in their own right, but they also provide ways to revise concepts at both levels of descriptions. In addition, the capacity of the framework to suggest predictions that can be empirically tested to uncover the components and processes responsible for certain cognitive behaviours makes the neurocomputational framework more progressive than the others in advancing our understanding of cognitive phenomena. If correct, these predictions would be informative (i.e. they would uncover something about the underlying mechanisms of which we were unaware) and also secure (i.e. they would depend on both psychophysical studies and on some features of neural circuits).

The thesis provides also a better understanding of some theoretical terms often adopted in cognitive science: cause and representation. Chapter 1 analyses the notion of cause within the context of causal explanations of cognitive behaviour and argues that causal statements cannot be grounded solely in counterfactual statements (e.g. Woodward, 2003, 2008). While counterfactuals are important epistemic tools, it is the identification and description of mechanisms that justify the existence of causes and causal relations. Chapter 3 examines the arguments used by advocates of the anti-representational framework against the usefulness of the notion of representation in explaining cognitive phenomena and shows that representations are required to explain a wide range of cognitive phenomena that do not result from a direct coupling between a system and its environment (i.e. representation-hungry problems). Chapter 4 discusses William Ramsey's attack to the notion of representation in connectionism and cognitive neuroscience (Ramsey, 2007). It argues that we are justified in treating a system as trafficking in representations when we explain its cognitive success in terms of internal models that the system employs to draw inferences about the world. These kinds of explanations are common within connectionism and cognitive neuroscience, hence certain internal neural states can be genuinely considered representational even in the absence of a full-blown theory of representation. Indeed, attributing representations to a system allows us to make its cognitive performance intelligible and predictable. Chapters 6 and 7 argue that the neurocomputational framework, thanks to its model-based methodology, allows the (empirical) exploration of various formats of representation. In particular, it leaves room for representational diversity and provides ways to deepen our understanding of the notion of representation in light of discoveries about the structure and functioning of the brain.

Good explanations of cognitive behaviour are, however, far from being simple. Given that models can explain to the extent that they incorporate ecological and biological considerations in their construction, two important challenges for neurocomputational models arise. First, if, as I suggested in chapters 6 and 7, neural systems carry out Bayesian and Reinforcement Learning algorithms to perform cognitively, then these algorithms must run quickly and efficiently. Yet, a system that implements Bayesian computations requires a significant amount of time and resources. This means that approximate forms of Bayesian computations have to be investigated and that new algorithms must be discovered. Second, for the framework to genuinely allow the identification of mechanisms underlying cognitive behaviour, it needs to provide researchers with a much clearer idea of what it means for specific neural circuits to realise an algorithm or some of its variables. The lack of a neat understanding of neural realisation negatively affects the goodness of the framework by limiting the effectiveness of the experiments and of the interpretations that scientists can make. On the positive side, the framework clearly strives for such clarification, as I have exemplified in my analysis of various cognitive behaviours in chapters 6 and 7. The top-down approach peculiar of the neurocomputational framework allows testable predictions, leaves room for exploring a broad range of different assumptions about how a cognitive system might perform a certain task and opens up the possibility for representational diversity. It also sheds new light on the possible bridges between functional and neural characterisations of behaviour: the personal level guides and motivates the search for mechanisms and provides important conceptual tools to understand the nature of cognitive phenomena, and the subpersonal level justifies the validity of personal-level claims through mechanisms, thus uncovering something of the constitutive nature of cognitive phenomena.

Bibliography

Anderson, J.R. (1990). *Rational Analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bechtel, W. (1998). "Representations and cognitive explanations: Assessing the dynamic challenge in cognitive science", *Cognitive Science*, vol.22, pp. 295–318.

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspective in cognitive neuroscience*. Routledge, London.

Bechtel, W., and Abrahamsen, A. (2005). "Explanation: A Mechanist Alternative". In C. Craver, and L. Darden (Eds.). *Special Issue: "Mechanisms in Biology" Studies in History and Philosophy of Biological and Biomedical Sciences*, vol.36, pp. 421–441.

Bechtel, W., and Richardson, R.C. (1993). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton, Princeton University Press.

Beck, J.M., Ma, W.J., Kiani, R., Hanks, T., Churchland, A.K., Roitman, J., Shadlen, M.N., Latham, P.E., and Pouget, A. (2008). "Probabilistic population coded for Bayesian decision making", *Neuron*, vol.60, pp. 1142–1152.

Bermudez, J.L. (2000). “Personal and Sub-personal: A Difference without a distinction”, *Philosophical Explorations*, vol.3, pp. 63–82.

Bermudez, J.L. (2003). *Thinking without words*. Oxford University Press.

Bermudez, J.L. (2005). *Philosophy of Psychology: a contemporary introduction*. Routledge, New York.

Bermudez, J.L. (2009). *Decision Theory and Rationality*. Oxford University Press.

Berniker, M, and Kording, K. (2008). “Estimating the sources of motor errors for adaptation and generalization”, *Nature Neuroscience*, vol.11(12), pp. 1454–1461.

Bickle, J. (2003). *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Dordrecht, Kluwer Academic Publishing.

Bickle, J. (2006). “Reducing mind to molecular pathways: explicating the reductionism implicit in current cellular and molecular neuroscience”, *Synthese*, vol.151, pp. 411–434.

Bickle, J. (2007). “Ruthless reductionism and social cognition”, *Journal of Physiology-Paris*, vol.101, pp. 230–235.

Blackemore, S.J., Wolpert, D.M., and Frith, C.D. (2002). “Abnormalities in the awareness of action”, *Trends in Cognitive Sciences*, vol.6(6), pp. 237–242.

Brooks, R. (1991). “Intelligence without representation”, *Artificial Intelligence*, vol.47, pp. 139–159.

Burge, T. (2010). *Origins of Objectivity*. Oxford University Press, New York.

Butts, D.A., and Goldman, M.S. (2006). “Tuning Curves, Neuronal Variability, and Sensory Coding”, *PLoS Biology*, vol.4(4):e92, pp. 0639–0646.

Chalk, M., Seitz, A.R., and Series, P. (2010). “Rapidly learned stimulus expectations alter perception of motion”, *Journal of Vision*, vol.10, pp. 1–18.

Chater, N., Oaksford, M., Nikisa, R., and Redington, M. (2003). "Fast, frugal, and rational: How rational norms explain behaviour", *Organizational Behaviour and Human Decision Processes*, vol.90, pp. 63–86.

Chemero, A. (2000). "Anti-representationalism and the Dynamical Stance", *Philosophy of Science*, vol.67(4), pp. 625–647.

Chemero, A. (2001). "Dynamic Explanations and Mental Representation", *Trends in Cognitive Science*, vol.5(4), pp.141–142.

Churchland, P. (2004). "How do neurons know", *Daedalus*, vol.133(1), pp. 42–50.

Churchland, P.M. (1981). "Eliminative materialism and the propositional attitudes", *The Journal of Philosophy*, vol.78(2), pp. 67–90.

Churchland, P.M. (1989). *A Neurocomputational Perspective*. MIT Press.

Churchland, P.M. (1998). "Conceptual Similarity across Sensory and Neural Diversity: the Fodor/Lepore Challenge Answered", *The Journal of Philosophy*, vol.XCV(1), pp. 5–32.

Churchland, P.S. (1986). *Neurophilosophy: toward a unified science of the mind-brain.* MIT Press.

Churchland, P.S. (1990). "Is neuroscience relevant to philosophy?", *Canadian Journal of Philosophy*, vol.16, pp. 323–341.

Clark, A. (1993). *Associative Engines*. MIT Press.

Clark, A. (1998). *Being there: putting brain, body and world together again*. MIT Press.

Clark, A. (2001). *Mindware: An Introduction to the Philosophy of Cognitive Science*. Oxford University Press.

Clark, A., and Toribo, J. (1994). "Doing Without Representing?", *Synthese*, vol.101, pp. 401–431.

Corlett, P.R., Honey, G.D., and Fletcher, P.C. (2007). “From prediction error to psychosis: ketamine as a pharmacological model of delusions”, *Journal of Psychopharmacology,* vol.21(3), pp. 238–252.

Corlett, P.R., Honey, G.D., Krystal, J.H., and Fletcher, P.C. (2011). “Glutamatergic Model Psychoses: Prediction Error, Learning, and Inference”, *Neuropsychopharmacology*, vol.36(1), pp. 294–315.

Corlett, P.R., Krystal, J.H., Taylor, J.R., and Fletcher, P.C. (2009). “Why do delusions persist?”, *Frontiers in Human Neuroscience*, vol.3, p. 12.

Craver, C. (2002). “Interlevel Experiments and multilevel mechanisms”, *Philosophy of Science*, vol.69, pp. S83–S97.

Craver, C. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Clarendon Press: Oxford.

Craver, C., and Darden, L. (2005). “Introduction: Mechanisms Then and Now”. In C. Craver, and L. Darden (Eds.). *Special Issue: “Mechanisms in Biology,” Studies in History and Philosophy of Biological and Biomedical Sciences*, vol.36, pp. 233–244.

Datteri, E., and Laudisa, F. (2012). “Model testing, prediction and experimental protocols in neuroscience: A case study”, *Studies in History and Philosophy of Biology and Biomedical Sciences*, vol.43, pp. 602–610.

Davidson, D. (1963). “Actions, Reasons and Causes”, *The Journal of Philosophy*, vol.XL(23), pp. 685–700.

Davidson, D. (1970). “Mental Events”, re-published in Davidson (1980). *Essays on Actions and Events.* pp. 207–227. Clarendon Press. Oxford.

Daw, N.D., Niv, Y., and Dayan, P. (2005). "Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioural control", *Nature Neuroscience*, vol.8(12), pp. 1704–1711.

Dayan, P., and Niv, Y. (2008). “Reinforcement Learning: The Good, The Bad and The Ugly”, *Current Opinion in Neurobiology*, vol.18, pp. 185–196.

De Jong, H.L., and Schouten, M.K.D. (2005). “Ruthless Reductionism: a review of John Bickle’s *Philosophy and Neuroscience: a ruthlessly reductive account*”, *Philosophical Psychology,* vol.18(4), pp. 437–486.

Dennett, D. (1969). *Content and Consciousness*. Routledge. London.

Doya, K., Ishii, S., Pouget, A., and Rao, R.P.N. (2007). *Bayesian Brain: Probabilistic Approaches to Neural Coding*. MIT Press.

Dretske, F. (1988). *Explaining Behavior: Reasons in A Word of Causes*. MIT press.

Earman, J., Roberts, J., and Smith, S. (2002). “Ceteris paribus lost”, *Erkenntnis*, vol.53(3), pp. 281–301.

Eliasmith, C. (2003). “Moving beyond metaphors: Understanding the mind for what itis”, *Journal of Philosophy*, vol.100(10), pp. 493–520.

Eliasmith, C. (2005). "A new perspective on representational problems", *Journal of Cognitive Science*, vol.6, pp. 97–123.

Ernst, M.O., and Banks, M.S. (2002). "Humans integrate visual and haptic information in a statistically optimal fashion", *Nature*, vol.415(24), pp. 429–433.

Evans, J., and Over, D.E. (1996). “Rationality in the selection task: Epistemic utility versus uncertainty reduction”, *Psychological Review*, vol.103, pp. 356–363.

Evans, J., and Over, D.E. (1996). *Rationality and Reasoning*, Psychology Press.

Evans, J., and Over, D.E. (1997). “Rationality in reasoning: The problem of deductive competence”, *Current Psychology of Cognition*, pp. 3–38.

Faure, A., Haberland, U., Conde', F., and Massioui, N.E. (2005). "Lesion to the nigrostriatal dopamine system disrupts stimulus-response habit formation", *Journal of Neuroscience*, vol.25(11), pp. 2771–2780.

Fletcher, P.C., and Frith, C.D. (2009). "Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia", *Nature Reviews Neuroscience*, vol.10(1), pp. 48–58.

Fodor, J. (1968). *Psychological Explanation*. New York, Random House.

Fodor, J. (1975). *The Language of Thought*. Harvard University Press.

Fodor, J. (1987). *The Problem of Meaning in the Philosophy of Mind*. MIT Press.

Fodor, J., and Pylyshyn, Z. (1988). "Connectionism and Cognitive Architecture: A Critical Analysis", *Cognition*, vol.28, pp. 3–71.

Friedman, M. (1974). "Explanation and Scientific Understanding", *The Journal of Philosophy*, vol.71(1), pp. 5–19.

Friston, K. (2005). "A theory of cortical responses", *Philosophical Transactions of the Royal Society London B Biological Sciences*, vol.360, pp. 815–836.

Friston, K. (2010). "The free-energy principle: a unified brain theory?", *Nature Reviews Neuroscience,* vol.11(2), pp. 127–138.

Friston, K., and Stephan, K. (2007). "Free energy and the brain", *Synthese*, vol.159(3), pp. 417–458.

Gazzaniga, M. (1988). *Perspectives in Memory Research*. MIT Press.

Gibson, J.J. (1979). *The Ecological Approach to Visual Perception.* Houghton Mifflin. Boston.

Gigerenzer, G. (2000). *Adaptive Thinking. Rationality in the Real World*. New York. Oxford University Press.

Gigerenzer, G., and Goldstein, D.G. (1996). “Reasoning the fast and frugal way: Models of bounded rationality”, *Psychological Review*, vol.103, pp. 650–669.

Gigerenzer, G., Todd, P.M., and the ABC Group (1999). *Simple heuristics that make us smart*. New York. Oxford University Press.

Giunti, M. (1995). “Dynamical Models of Cognition”. In R. Port, and T. van Gelder (Eds.). *Mind as Motion: Explorations in the dynamics of cognition*. pp. 549–571, MIT Press.

Glennan, S. (1996). “Mechanisms and the Nature of Causation”, *Erkenntnis*, vol.44, pp. 49–71.

Glennan, S. (2005). “Modeling Mechanisms”. In C. Craver and L. Darden (Eds.). Special Issue: “Mechanisms in Biology,” *Studies in History and Philosophy of Biological and Biomedical Sciences*, vol.36, pp. 443–464.

Glennan, S. (2010). "Mechanisms, causes and the layered model of the world", *Philosophy and Phenomenological Research*, vol.81, pp. 362–381.

Gold, I., and Stoljar, D. (1999). “A neuron doctrine in the philosophy of neuroscience”, *Behavioural and Brain Sciences*, vol.22, pp. 809–869.

Gould, S.J., and Gould, C.J. (1998). “Reasoning in animals”, *Scientific American*, vol.9, pp. 52–59.

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*, vol.14(8), pp. 357–364.

Grush, R. (2003). “In defence of some 'Cartesian' assumptions concerning the brain and its operation”, *Biology and Philosophy*, vol.18, pp. 53–93.

Grush, R. (2004). “The emulation theory of representation: motor control, imagery, and perception”, *Behavioural and Brain Sciences*, vol.27, pp. 377–442.

Haken, H., Kelso, J.A.S., and Bunz, H. (1985). “A theoretical model of phase transitions in human hand movements", *Biological Cybernetics*, vol.51, pp. 347–356.

Hawkins, R.D., and Kandel, E.R. (1984). "Is there a cell-biological alphabet for simple forms of learning?", *Psychological Review*, vol.91, pp. 376–391.

Hempel, C. (1965). *Aspects of Scientific Explanation*. New York. Free Press.

Hirstein, W. (2005). *Brain Fiction: Self-deception and the Riddle of Confabulation*. MIT Press.

Hohwy, J. (2004). "Top-down and bottom-up in delusion formation", *Philosophy, Psychiatry and Psychology*, vol.11(1), pp. 65–70.

Hohwy, J. (2012). "Attention and conscious perception in the hypothesis testing brain", *Frontiers in Psychology*, vol.3(96), pp. 1–14.

Hohwy, J. (2013). "Delusions, illusions, and inference under uncertainty", *Mind & Language*, vol.28, pp. 57–71.

Hohwy, J., and Rosenberg, R. (2005). "Unusual experiences, reality testing and delusions of alien control", *Mind & Language*, vol.20(2), pp. 141–162.

Hohwy, J., Roepstorff, A., and Friston, K. (2008). "Predictive coding explains binocular rivalry: an epistemological review", *Cognition*, vol.108(3), pp. 687–701.

Hornsby, J. (2000). "Personal and subpersonal. A defence of Dennett's early early distinction", *Philosophical Explorations*, vol.3, pp. 6–24.

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, and D.G. Beiser (Eds.). *Models of information processing in the basal ganglia*. pp. 249–270, MIT Press.

Hubel, D.H., and Wiesel, T.N. (1962). "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex", *The Journal of Physiology*, vol.160, pp. 106–154.

Jeannerod, M. (1995). "Mental Imagery in The Motor Context", *Neuropsychologia*, vol.11, pp. 1419–1432.

Kandel, E.R., and Schwartz, J.H. (1982). “Molecular biology of learning: modulation of neurotransmitter release”, *Science*, vol.218(4571), pp. 433–443.

Kaplan, D.M., and Bechtel, W. (2011). “Dynamical Models: An Alternative or Complement to Mechanistic Explanations”, *Topics in Cognitive Science*, vol.3, pp. 438–444.

Kapur, S. (2003). “Psychosis as a State of Aberrant Salience: A Framework Linking Biology, Phenomenology, and Pharmacology in Schizophrenia”, *The American Journal of Psychiatry*, vol.160, pp. 13–23.

Keijzer, F. (1998). "Doing without representations which specify what to do", *Philosophical Psychology*, vol.11(3), pp. 269–302.

Keijzer, F. (2005). "Theoretical Behaviourism Meets Embodied Cognition: Two Theoretical Analysis of Behaviour", *Philosophical Psychology*, vol.18(1), pp. 128–143.

Kelso, J.A.S. (1995). *Dynamic Patterns: The Self Organization of Brain and Behaviour*. MIT Press.

Kemp, C., and Tenenbaum, J.B. (2008). “The discovery of structural form”, *Proceedings of the National Academy of Science*, vol.105(31), pp. 10687–10692.

Kitcher, P. (1981). “Explanatory Unification”, *Philosophy of Science*, vol.48, pp. 507–531.

Knill, D.C., and Pouget, A. (2004). "The Bayesian brain: the role of uncertainty in neural coding and computation", *Trends in Cognitive Science*, vol.27(12), pp. 712–719.

Knill, D.C., and Richards, W. (1996). *Perception as Bayesian Inference.* New York. Cambridge University Press.

Knill, D.C., Kersten, D., and Yuille, A. (1996). “A Bayesian formulation of visual perception”. In D.C. Knill, and W. Richards (Eds.). *Perception as Bayesian Inference*. pp. 1–21. Cambridge University Press.

Kogan, J.H., Franklandand, P.W., and Silva, A.J. (2000). “Long-term memory underlying hippocampus-dependent social recognition in mice”, *Hippocampus*, vol.10(1), pp. 47–56.

Kording, K.P., Beierholm, U., Ma, W.J., Quartz, S., Tenenbaum, J.B., and Shams, L. (2007). "Causal Inference in Multisensory Perception", *PLoS ONE*, vol.2(9):e943, pp. 1–10.

Lettvin, J.Y., Maturana, H.R., McCulloch, W.S., and Pitts, W.H. (1959). “What the frog’s eye tells the frog’s brain”, *Journal of the Institute of Radio Engineers*, vol.47, pp. 1940–1951.

Lewis, D. (1973). “Causation”, *The Journal of Philosophy*, vol.70(17), pp. 556–567.

Ma, W.J., Beck, J.M., Lathan, P.E., and Pouget, A. (2006). "Bayesian inference with probabilistic population codes", *Nature Neuroscience*, vol.9(11), pp. 1432–1438.

Machamer, P. (2004). “Activities and Causation: The Metaphysics and Epistemology of Mechanisms”, *International Studies in the Philosophy of Science*, vol.18, pp. 27–39.

Maloney, L. T., and Mamassian, P. (2009). “Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer”, *Visual Neuroscience*, vol.26, pp. 147–155.

Marr, D. (1982). *Vision: A Computational Investigation in to the Human Representation and Processing of Visual Information*. Freeman. New York.

Mars, R.B., Shea, N., Kolling, N., and Rushworth, M.F.S. (2012). "Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control", *Quarterly Journal of Experimental Psychology*, vol.65(2), pp. 252–267.

McCauley, R.N., and Bechtel, W. (2001). “Explanatory Pluralism and Heuristic Identity Theory”, *Theory Psychology*, vol.11(6), pp. 736–760.

McDowell, J. (1994). “The content of perceptual experience”, *The Philosophical Quarterly*, vol.44(175), pp. 190–205.

Mercier, H., and Sperber, D. (2011). “Why do humans reason? Arguments for an argumentative theory”, *Behavioural and Brain Sciences*, vol.34, pp. 57–111.

Millikan, R. (1984). *Language, Thought and other Biological Categories*. MIT Press.

Montague, P.R., Hyman, S.E., and Cohen, J.D. (2004). "Computational roles for dopamine in behavioural control", *Nature Publishing Group*, vol.431, pp. 760–767.

Mukamel, R., Gelbard, H., Arieli, A., Hasson, U., Fried, I., and Malach, R. (2005). “Coupling between neuronal firing, field potentials, and fMRI in human auditory cortex”, *Science*, vol.309, pp. 951–954.

Musallam, S., Corneil, B., Greger, B., Scherberger, H., and Andersen, R. (2004). “Cognitive control signals for neural prostethics”, *Science*, vol.305, pp. 258–262.

Nisbett, R.E., and Wilson, T.D. (1977). “Telling more than we can know: Verbal reports on mental processes”, *Psychological Review*, vol.84(3), pp. 231–259.

Nisbett, R.E., and Wilson, T.D. (1978). “The accuracy of verbal reports about the effects of stimuli on evaluations and behavior”, *Social Psychology*, vol.41(2), pp. 118–131.

Niv, Y., and Schoenbaum, G. (2008). "Dialogues on prediction error", *Trends in Cognitive Sciences*, vol.12(7), pp. 265–272.

O'Keefe, J., and Conway, D.H. (1978). "Hippocampal place units in the freely moving rat: Why they fire when they fire", *Experimental Brain Research*, vol.31(4), pp. 573–590.

Pellicano, E., and Burr, D. (2012). "When the world becomes too real: a Bayesian explanation of autistic perception", *Trends in Cognitive Sciences*, vol.16, pp. 504–510.

Piccinini, G. (2007). "Computational Explanation and Mechanistic Explanation of Mind". In M. de Caro, F. Ferretti, and M. Marraffa (Eds.). *Cartographies of the Mind: The Interface between Philosophy and Cognitive Science*. pp. 23–36, Dordrecht: Springer.

Piccinini, G., and Craver, C. (2011). "Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches", *Synthese*, vol.183(3), pp. 283–311.

Port, R., and van Gelder, T. (1995). *Mind as Motion: Explorations in the dynamics of cognition*. MIT Press.

Psillos, S. (2004). "A Glimpse of the Secret Connexion: Harmonizing Mechanisms with Counterfactuals", *Perspectives on Science*, vol.12, pp. 288–319.

Psillos, S. (2007). "Causal Explanation and Manipulation". In J. Person, and P. Ylikoski. *Rethinking Explanation.* pp. 97–112, Boston Studies in the Philosophy of Science, Springer.

Ramsey, W. (2007). *Representation Reconsidered*. Cambridge University Press.

Ramsey, W., and Stich, S. (1990). "Connectionism, Eliminativism and the Future of Folk Psychology". *Philosophical Perspectives*, vol.4, pp. 499–533.

Rao, R.P.N., and Ballard, D.H. (1999). "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive field effects", *Nature Neuroscience*, vol.2(1), pp. 79–87.

Rao, R.P.N., Olshausen, B., and Lewicki, M. (2002). *Probabilistic Models of the Brain: Perception and Neural Function*. MIT Press.

Rescorla, R.A. (1968). “Probability of shock in the presence and absence of CS in fear conditioning”, *Journal of Comparative and Physiological Psychology*, vol.66(1), pp. 1–5.

Rescorla, R.A., and Skucy, J.C. (1969). “Effect of Response-Independent Reinforcers During Extinction”, *Journal of Comparative and Physiological Psychology*, vol.67, pp. 381–389.

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press.

Schultz, W., Dayan, P., and Montague, P.R. (1997). “A neural substrate of prediction and reward”, *Science*, vol.275(5306), pp. 1593–1599.

Seidenberg, M., and McClelland, J. (1989). “A Distributed Developmental Model of Word Recognition and Naming”, *Psychological Review*, vol.96, pp. 523–568.

Shagrir, O. (2012). “Structural Representations and the Brain”, *British Journal of Philosophy of Science*, vol.0, pp. 1–27.

Shea, N. (2007). “Content and Its Vehicles in Connectionist Systems”, *Mind & Language*, vol.22(3), pp. 246–269.

Shea, N. (2013). "Naturalising representational content", *Philosophy Compass*, vol.8(5), pp. 496–509.

Shea, N. (*forthcoming*), “Neural mechanisms of decision-making and the personal level”. In K.W.M. Fulford, M. Davies, G. Graham, J. Sadler, G. Stanghellini, and T. Thornton (Eds.). *Oxford Handbook of Philosophy and Psychiatry*. Oxford University Press.

Smolensky, P. (1988). "On the Proper Treatment of Connectionism", *Behavioural and Brain Sciences*, vol.11, pp. 1–74.

Snow, N. (2006). "Habitual virtuous actions and automaticity", *Ethical Theory and Moral Practice*, vol.9(5), pp. 545–561.

Sprevak, M. (2011). "William M. Ramsey, *Representations Reconsidered*", *British Journal of Philosophy of Science*, vol.62, pp. 669–675.

Stich, S., and Nichols, S. (2003). "Folk Psychology". In S. Stich, and T.A. Warfield (Eds.). *The Oxford Guide to Philosophy of Mind*. pp. 235–255, Basil Blackwell, Oxford.

Sullivan, J.A. (2009). "The multiplicity of experimental protocols: a challenge to reductionist and non–reductionist models of the unity of neuroscience", *Synthese*, vol.167, pp. 511–539.

Swenson, R., and Turvey, M.T. (1991). "Thermodynamic reasons for perception-action cycles", *Ecological Psychology*, vol.3, pp. 317–348.

Tenenbaum, J.B., Kemp, C., Griffiths, T.L., and Goodman, N.D. (2011). "How to grow a mind: statistics, structure, and abstraction", *Science*, vol.331, pp. 1279–1285.

Turvey, M. T., and Carello, C. (1981). "Cognition: The view from ecological realism", *Cognition*, vol.10, pp. 313–321.

van Gelder, T. (1995). "What might cognition be, if not computation?", *Journal of Philosophy*, vol.91, pp. 345–381.

van Gelder, T. (1998). "The Dynamical Hypothesis in Cognitive Science", *Behavioural and Brain Sciences*, vol.21, pp. 1–12.

Weiss, Y., Simoncelli, E.I., and Adelson, E.H. (2002). "Motion illusions as optimal percepts", *Nature Neuroscience*, vol.5, pp. 598–604.

Schultz, W. (2010). "Dopamine signals for reward value and risk: basic and recent data", *Behavioural and Brain Functions*, vol.6(24), pp. 1–9.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation.* Oxford University Press.

Woodward, J. (2008). "Mental Causation and Neural Mechanism". In J. Hohwy, and J. Kallestrup (Eds.). *Being Reduced: New Essays on Reduction, Explanation, and Causation*. pp. 218–262, Oxford University Press.

Yin, H.H., Knowlton, B.J., and Balleine, B.W. (2004). "Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning", *European Journal of Neuroscience*, vol.19, pp. 181–189.