Towards Explanatory Pluralism in Cognitive Science

Maria Serban

This dissertation is submitted for the degree of
Doctor in Philosophy

University of East Anglia
School of Philosophy
February 2014

ABSTRACT

This thesis seeks to shed light on the intricate relationships holding between the various explanatory frameworks currently used within cognitive science. The driving question of this philosophical investigation concerns the nature and structure of cognitive explanation. More specifically, I attempt to clarify whether the sort of scientific explanations proposed for various cognitive phenomena at different levels of analysis or abstraction differ in significant ways from the explanations offered in other areas of scientific inquiry, such as biology, chemistry, or even physics. Thus, what I will call *the problem of cognitive explanation*, asks whether there is a distinctive feature that characterises cognitive explanations and distinguishes them from the explanatory schemas utilised in other scientific domains.

I argue that the explanatory pluralism encountered within the daily practice of cognitive scientists has an essential *normative* dimension. The task of this thesis is to demonstrate that pluralism is an appropriate standard for the general explanatory project associated with cognitive science, which further implies defending and promoting the development of multiple explanatory schemas in the empirical study of cognitive phenomena.

CONTENTS

1

INTRODUCTION

1.1 THE PROBLEM

This thesis seeks to shed light on the intricate relationships holding between the various explanatory frameworks currently used within cognitive science. The driving question of this philosophical investigation concerns the nature and structure of cognitive explanation. More specifically, I attempt to clarify whether the sort of scientific explanations proposed for various cognitive phenomena at different levels of analysis or abstraction differ in significant ways from the explanations offered in other areas of scientific inquiry, such as biology, chemistry, or even physics. Thus, what I will call *the problem of cognitive explanation*, asks whether there is a distinctive feature that characterises cognitive explanations and distinguishes them from the explanatory schemas utilised in other scientific domains.

The philosophical project of analysing the notion of cognitive explanation confronts two important challenges. The current landscape of cognitive science displays a multiplicity of theoretical and experimental frameworks developed in order to deal with a wide variety of cognitive problems. A philosophical account which attempts to make sense of the actual situation in scientific practice needs to acknowledge this variety and develop the critical tools for evaluating the strengths and limitations of each available approach. In other words, an appropriate philosophical treatment of the problem of cognitive explanation should reflect the diversity of the explanatory schemas utilised by practicing cognitive scientists. Beside satisfying this *descriptive adequacy* requirement, a philosophical account of cognitive explanation should also say what distinguishes cognitive explanation from other scientific achievements obtained in the investigation of cognitive phenomena. Thus, the second type of challenge facing this project can be characterised as a search for the common features of cognitive explanations.

This thesis attempts to show that the explanatory pluralism encountered within the daily practice of cognitive scientists has an essential *normative* dimension: the task is to demonstrate that pluralism is an appropriate standard for the general explanatory project associated with cognitive science, which further implies defending and promoting the development of multiple explanatory schemas in the empiri-

cal study of cognitive phenomena. The difficulty of this task derives in part from the fact that the two challenges sketched above seem to pull the account in two opposing directions. On the one hand, there is the requirement to recognise the *de facto* plurality of explanatory strategies used in the study of cognitive phenomena. On the other hand, there is the strong intuition that there must be something distinctive about (good) scientific explanations of cognitive phenomena.

An apparently easy way to solve this tension would be to claim that this form of pluralism is merely a faithful representation of the state of art in a very young and immature science. That is, one is encouraged to tolerate explanatory pluralism at present because it is a way of progressing towards a more stable and mature cognitive science, within which one would be able to identify the dominant and correct notion of cognitive explanation. Instead, I propose a different way to solve the tension created by the descriptive adequacy and normative challenges by putting forward a substantive argument in support of the normative character of explanatory pluralism.

1.2 EXPLANATORY VS. ONTOLOGICAL CONCERNS

The project of analysing the structure of cognitive explanation differs in significant respects from another type of concern that has been traditionally at the forefront of the field of philosophy of mind, namely the metaphysical or ontological concern with the kind of 'stuff' the mind must be made of. Opposing very general positions ranging from materialism (the view that all that exists is matter) to one or another form of dualism (the view that mental things exist over and above material things), traditional ontological debates focus on highly abstract questions which seem to be only loosely connected to the actual scientific practices of cognitive scientists. The strategy advocated in this thesis attempts to show how specific ontological commitments arise and become established as essential ingredients in the epistemic activities of practising scientists. This perspective also promises to elucidate the roles that ontological principles/commitments play in the construction and refinement of scientific explanations of cognitive phenomena.

By considering ontological questions within a broadly methodological framework, one is in a better position to say how such considerations actually guide and constrain particular scientific activities which aim to advance our understanding of cognitive phenomena. That is, a methodological approach allows one to show that, in order to engage in a specific epistemic activity, one must assume the truth of some particular metaphysical or ontological principles. These principles can be taken to guarantee the intelligibility and performability of the scientific practices encountered in an area of empirical inquiry such as cognitive science. However, unlike more traditional philosophi-

cal treatments of these issues, the proposed account does not entail that ontological principles are universal or necessarily true in some absolute sense. Rather than seeking to establish their precise modal strength, the account acknowledges that there are different metaphysical principles which intervene in certain scientific activities such as measurement, prediction, confirmation, and explanation.

More specifically, these investigations aim to pin down the roles played by specific ontological principles in the practice of constructing adequate potential explanations of particular cognitive phenomena. As will be shown in the following chapters, different philosophical models of cognitive explanation identify different ontological principles that partly determine what counts as a good explanation of a given cognitive phenomenon. Some of these models highlight very specific ontological commitments such as the existence of particular neurobiological mechanisms and their component entities, while others appeal to general principles such as the principle of discreteness, transitivity, single value, causality, and the principle uniform consequence, among others.

One important advantage of locating meaningful ontological questions in a methodological setting in which one focuses on the various explanatory schemas used in cognitive science is that this strategy avoids some of the pitfalls of traditional metaphysical debates concerning the nature of the mind. On the one hand, positions such as materialism (physicalism), dualism or hybrid versions of non-reductive physicalism or anomalous materialism risk, because of their generality, remaining utterly uninstructive when it comes to the evaluation of specific hypotheses concerning the structure and nature of cognitive capacities. None of these positions seems to be adequately equipped to settle questions about the actual structure and organisation of particular cognitive capacities, such as whether conceptual structures impact object recognition or the early visual system provides sufficiently well-structured information to guide the recognition task, whether language acquisition is a generalised statistical process or is governed by a set of innate rules, and so on. On the other hand, the prioritisation of ontological debates in the cognitive domain has tended to distort the dynamics and aims of cognitive scientific research by encouraging the petrification of some of the concepts used to account for different cognitive phenomena at certain levels of analysis or abstraction. The danger of this latter tendency is that it rules out, on purely *a priori* grounds, certain research strategies and explanatory schemas as being inappropriate for the study of cognitive phenomena.

Thus, rather than seeking to defend and refine any of these general positions, I propose to look at ontological principles as scientifically operative in that they contribute to the construction of better scientific explanations which in turn advance our understanding of real-

ity. The insights made available by this perspective are compatible with a moderate form of scientific realism (cf. Dupré 1993; Chang 2009, 2012). The operational notion of reality (or external world) involved in this conception does not imply that our theories give us a direct representation of the external world. In fact, by focusing on the various explanatory schemas used in different branches of cognitive science, one is better placed to appreciate that the limitations of our understanding of cognitive phenomena are partly due to the complexity of the external world and its resistance to conforming to our well-behaved categories.

1.3 PLAN AND ARGUMENT OF THE THESIS

As outlined above, the principal aim of the following investigations is to offer a substantial account of the notion of cognitive explanation which successfully balances the diversity of explanatory frameworks currently used in the study of cognition and the need of a more precise characterisation of the notion of explanatory value itself. That is, an account which is both descriptively adequate with respect to the multiplicity of explanatory schemas used in cognitive science, and normative in the sense of providing a way of evaluating the relative explanatory power/value of the various accounts proposed in this domain of empirical inquiry. The argument strategy adopted for this purpose combines two distinct, yet interrelated, perspectives. First, I analyse the general principles and assumptions that underlie a series of important philosophical accounts of the notion of cognitive explanation. Second, I adopt a practice-based perspective and consider how these principles relate to the actual activities of constructing and evaluating explanations of different cognitive phenomena. Together, these analyses will provide important guidelines for the articulation of a novel and more adequate account of cognitive explanation.

This style of argumentation reflects two important constraints on the type of philosophical analysis that I take to be appropriate for tackling the problem of cognitive explanation. First, the account I put forward is rooted in the careful analysis of several paradigmatic explanatory models of cognitive capacities. This reverses the order of more traditional philosophical analyses which tend to identify a set of highly general principles and then seek to show how these apply to actual cases of cognitive explanations. However, this proposal does not merely aim to analyse and describe (faithfully) the actual practice of constructing explanations of particular cognitive phenomena, but also to use these analyses to articulate a more general view of what counts as a cognitive explanation in the first place. The emerging view will have a distinctive normative component that will not be justifiable solely in virtue of the descriptive adequacy of the account, but also by appealing to the role that it plays in guiding the further

investigation of cognitive phenomena and the development of better cognitive explanations.

Second, the normative character of the proposed account will have to cover several levels of analysis. At the most general level, it will have to specify what counts as an explanatory structure in the first place. In addition, the account must clarify whether there are any specific norms which govern the construction of scientific explanations of *cognitive* phenomena. And, finally, it will have to investigate the more fine-grained norms or principles which guide the construction of specific models/theories of cognitive capacities. For the latter part of the investigation, the account will have to pay special attention to the details and individual differences of the cognitive phenomena targeted by different types of explanation.

At this point, the difficulty of formulating a general and normative account of explanation arises again precisely because one hopes to do justice to a wide range of explanatory schemas used across the various sub-branches of cognitive science. A nice feature of thinking about the aims of a philosophical account of cognitive explanation along these lines is that it respects the intuition that the best way of approximating it is to have two camps, one pressing for the uniformity and generality of the notion of explanation and the other insisting on the details of particular modes/styles of explanation, each camp contributing to the validation of the insights made available by the other.

Relying on this general strategy, the rest of the thesis is organised in six different chapters that analyse the problem of cognitive explanation from a series of complementary perspectives. Chapter 2 revisits some of the most prominent philosophical accounts of the notion of scientific explanation. Starting with the classical covering law model of scientific explanation (Hempel and Oppenheim 1948; Hempel 1965), I then go on to discuss the statistical/probabilistic and causal accounts, which have emerged as critical responses to the Hempelian account. Whilst the original Hempelian model emphasises the inferential structure of explanation and the essential role played by natural laws in the construction of (good) scientific explanations, the probabilistic and causal accounts identify the source of the explanatory value of scientific theories in their capacity: (i) to reveal salient probabilistic patterns in the phenomena being investigated (Jeffrey 1969; Salmon 1971; Mellor 1976) and (ii) to show how the target phenomena fit into the causal structure of the world (Salmon 1984a, 1989). The main rationale for focusing on these three philosophical models of scientific explanation is that they have shaped in significant ways the landscape of current models of cognitive explanation. In addition, I briefly discuss the relation between these conceptions and the idea that explanatory theories/models have a distinctive unificatory power (Friedman 1974; Kitcher 1989). Finally,

I address the question of how one should seek to account for the pragmatic dimension of scientific explanation (van Fraassen 1980) in an analysis that attempts to provide a general characterisation of the structure of scientific explanation. The primary aim of these preliminary analyses is to provide a series of insights that would guide the development of an adequate conception of cognitive explanation.

Against this conceptual background, I then introduce in more detail the problem of cognitive explanation. Essentially, the issue consists in examining whether the explanatory schemas proposed in the various sub-branches of cognitive science constitute a distinct subspecies of scientific explanation. In setting up this problem, I point out that the orthodox view embraced by most philosophers of mind maintains that none of the existing philosophical models of scientific explanation provides an adequate picture of the explanatory aims and strategies used by practising psychologists and cognitive scientists (cf. Cummins 2010). This general way of thinking about cognitive explanation has in turn created a strong artificial dichotomy between the methodology of naturalistic fields of scientific inquiry, such as biology, chemistry or physics, and the methodology of psychology or cognitive science (cf. Wilson 1985). Furthermore, various forms of this dichotomy have been used as launchpads for arguing that the styles of reasoning and explanation which are adequate with regard to the proprietary objects of cognitive science are relatively independent from those deployed in other areas of scientific inquiry. In particular, the idea that functional or interpretative analysis constitutes the distinctive mark of psychological explanations arises from this particular way of conceiving the nature of cognitive explanation as essentially distinct from other types of scientific explanation (e.g., Fodor 1974; Cummins 1983, 2010).

Each of the following four chapters takes a prominent model of cognitive explanation and analyses its main strengths and weaknesses from a conceptual and empirical practice-based perspective. I argue, in particular, that the adoption of the latter viewpoint is essential for appreciating the limitations of the widespread commitment for *explanatory monism*, according to which all scientific theories/models must conform to the same standard of explanatory 'goodness'. In addition, I claim that the practice-based perspective allows the articulation of a novel and more adequate conception of cognitive explanation which acknowledges the various insights made available by existing models of cognitive explanation. Although the collection of models of cognitive explanation considered in these chapters is not meant to exhaust the variety of explanatory strategies used in all the different areas of cognitive research, I do take it to comprise a representative sample which will identify a number of the issues that other similar accounts of cognitive explanation are likely to face. Moreover, the type of philosophical analysis that I promote with respect to

the problem of cognitive explanation has much to gain from a close inspection of a few paradigmatic examples drawn from the current modelling and theorising practices of cognitive scientists. By paying closer attention to the individual problems addressed by particular cognitive theories/models, one is more likely to develop a critical outlook of the difficulties that confront the search for better theories of cognition, as well as the exciting (and sometimes unexpected) advancements made in understanding complex cognitive phenomena.

In line with this general agenda, chapter 3 explores the *mechanistic view of cognitive explanation*. The main reason for starting with the mechanistic conception is that the view promises to offer a much needed reconciliation between traditional accounts of scientific explanation (outlined in chapter 1) and the hypothesis that cognitive explanations have a distinctive character (e.g., Woodward 2000; Craver 2007b; Bechtel 2008). By exploring the main arguments put forward by various philosophers and cognitive scientists in defence of the mechanistic view, I seek to identify the main tenets of this model of cognitive explanation. In particular, I focus on two major theses associated with the mechanistic conception of explanation, namely that: (i) mechanistic explanation is a form of decompositional and constitutive explanation, and (ii) the view provides a general framework for integrating all the various explanatory strategies pursued in the different sub-branches of cognitive science. I argue against the monist assumption implicit in the latter hypothesis, and show that it leads to a number of difficult problems for the general mechanistic position. Overall, I aim to modify the mechanist view in such a way that its scope is clearly delimited and its significance within these constraints is correctly appreciated.

In chapter 4, I turn to the analysis of *classical computationalism*, according to which cognitive phenomena can be explained by postulating mental (internal) symbols and operations appropriately defined over them (e.g., Fodor 1980; Pylyshyn 1984; Gallistel and King 2009). The strategy proposed in this chapter distinguishes between two issues which are usually conflated in discussions of classical computationalism and its associate hypothesis concerning the structure of cognitive explanation. These are the computational individuation issue and the computational explanation issue. I offer a number of compelling reasons in support of the adoption of this argument strategy, claiming that it contributes to: (i) the clarification of the main theoretical principles of classical computational approaches to cognition, and (ii) the derivation of a better strategy for assessing the explanatory value of classical computational models of specific cognitive capacities.

Chapter 4 also urges an important shift of emphasis away from the more traditional metaphysical debates that have flourished around the classical computationalist thesis towards a more informed anal-

ysis of the modelling and theorising practices that rely on the theoretical principles advocated by classical computationalists. Drawing on the critical analysis of some of the most important arguments discussed in the scientific and philosophical literature (e.g., Fodor 1975; Marr 1982; Stich 1983; Cummins 1989; Pylyshyn 1984; Gallistel and King 2009; Egan 1992, 1999, 2010), I argue for the compatibility between a structural (internalist) computational *individuation* thesis and a quasi-pragmatic (externalist) thesis concerning the role of mental contents in classical computational *explanations* of specific cognitive capacities. In addition, I support the conceptual arguments put forward for this double thesis with a series of examples drawn from the recent computational modelling literature.

In chapter 5, I consider another version of the mechanistic conception of cognitive explanation that promises to constitute a substantial improvement on classical computationalist approaches to cognition. This form of the mechanistic view characterises computational explanations of specific cognitive capacities as a sub-species of mechanistic explanations (cf. Piccinini 2007b; Craver and Piccinini 2011). Furthermore, the view argues for a wide functional individuation strategy of the computational states and structures postulated by computational models/theories of cognition. I review the main arguments put forward in support of this position, seeking to establish whether it constitutes a radical alternative to classical views of computational explanations, an extension of such accounts, or is merely complementary. As in the previous chapter, I emphasise the separability of the computational individuation and computational explanation issues, and show that the mechanistic conception of computational explanation risks conflating the two, thereby blurring the criteria for evaluating good computational models/theories of cognitive phenomena.

Special attention is dedicated to the claim that the *mechanistic conception of computational explanation* provides a better strategy for dealing with the so-called realisation problem (Piccinini 2008a; Piccinini and Bahar 2013). The realisation problem points out that there is a significant disconnect between abstract (often functional) characterisations of cognitive phenomena and the neurobiological descriptions of the mechanisms that support or otherwise maintain cognitive processing. Because the realisation problem has been taken to constitute a cornerstone of any candidate theory/model of cognitive phenomena, the claim that it can be appropriately addressed within a broadly mechanistic framework seems to confer a definite advantage on the mechanistic model of computational explanation, since proponents of the mechanistic view tend to argue that mechanistic models of cognitive phenomena are biologically more plausible than the abstract models discussed by classical computationalists. In response, I argue that the realisation problem is still too poorly understood to be used as a reliable criterion for distinguishing between explanatory

and non-explanatory theoretical approaches to cognition. In addition, I point out that focusing on the realisation problem tends to obscure the explanatory structure of what are currently considered successful explanatory mechanistic models of cognition.

In chapter 6, I turn to the *connectionist conception of cognitive explanation*, which is also often portrayed as a radical alternative to the classical view of computational explanation. I analyse the connectionist position from two distinct but interrelated perspectives. At the conceptual level, I investigate what distinguishes connectionist from classical computation, and ask whether the two frameworks are indeed committed to different computational individuation strategies or not. Following these theoretical investigations, I adopt a practice-based standpoint in order to determine the distinctive features of connectionist explanations of cognition. I point out that, when it is constructed as a completely new and radical explanatory strategy for the cognitive domain, connectionism runs the risk of deflating entirely the notion of cognitive explanation. Another important feature of the connectionist framework consists in its principled and pragmatic commitment to mechanism. This raises further issues about the explanatory hypothesis associated with connectionism, and how it differs from the explanatory strategies analysed in the previous chapters of the thesis.

Finally, chapter 7 draws together the main lessons afforded by the investigation of these different philosophical models, and articulates a more systematic picture of cognitive explanation. In addition, I examine the consequences of adopting the proposed picture of scientific explanation for two salient problems that surface quite often in the philosophy of cognitive science: the problem of integrating (or unifying) accounts proposed in different branches of cognitive science, and the scientific realism issue regarding the entities postulated in the context of cognitive scientific research. The latter set of considerations reflects the fact that the novel approach to the problem of cognitive explanation advocated in this thesis yields important insights regarding distinct problems which arise in both philosophy of science and philosophy of mind. I conclude by assessing the promise and limitations of the account presented in the thesis and clarify further the implications of adopting the proposed version of explanatory pluralism within and outside the domain of cognitive science.

1.4 SCOPE AND FOCUS

This thesis advocates an important shift of perspective from traditional metaphysical or ontological debates concerning the nature and structure of the *mind* towards a detailed analysis of various types of explanatory accounts of specific mental/cognitive phenomena. This reorientation reflects the fact that more and more debates within phi-

losophy of mind tend to be informed by the most recent hypotheses put forward in the various sub-fields of cognitive science.

At the most general level, the distinctive contribution of this thesis consists in the overt questioning of the relationship between these two fields of philosophical investigation (i.e., philosophy of mind and philosophy of science) and the original integration of the analytic tools each of them makes available for the investigation of particular cognitive problems. By drawing on resources from both philosophical fields, I seek to derive a novel and adequate conception of the notion of cognitive explanation. At a more 'local' level, I derive a number of important consequences that bear on certain specific topics discussed in the philosophical literature, such as the question concerning what count as the constitutive principles of classical computationalism, mechanism, and connectionism, and their applicability to the study of cognitive phenomena. Due to its focus on the notion of cognitive explanation, this thesis does not analyse in a systematic way any specific empirical hypotheses/theories of particular cognitive capacities. Rather, the models discussed in the different chapters of the thesis are intended to illustrate the variety of cognitive problems investigated by practising scientists (in vision, memory, and language studies). This strategy is also intended to provide additional motivation for the version of explanatory pluralism advocated in the thesis.

Pluralism in science is a widely recognised 'fact' which reflects the multiplicity of models, theoretical approaches, and explanations encountered in different areas of scientific inquiry. Pluralism has also been defended from various perspectives in the philosophical literature (e.g., Dupré 1993; Cartwright 1999; Mitchell 2003; Chang 2012), and more particularly in the philosophy of psychology and cognitive science (e.g., Chemero and Silberstein 2008; Dale, Dietrich, and Chemero 2009; Stepp, Chemero, and Turvey 2011). Thus, whilst the moderate pluralistic view I defend here is not a particularly new or radical claim, it does make, or so I shall argue, an important contribution to current debates in the philosophy of psychology and cognitive science. I aim to offer additional motivation for endorsing explanatory pluralism not only as a descriptively adequate account of the diversity of views found in contemporary science, but also as a broadly normative thesis that should guide further philosophical analysis of the actual situation in cognitive scientific practice. Also, by drawing on the critical analyses of the philosophical models discussed in chapters 3 to 6, I seek to articulate a more precise account of the various explanatory structures utilised in the different disciplines of cognitive science.

Having argued from both a principled and practice-based perspective for a substantial pluralistic account of cognitive explanation, I claim that there are good reasons for attempting to extend the proposed view of explanation to other fields of scientific inquiry, al-

though, for reasons of space, I will have time only to raise this issue and offer a preliminary exploration of its supporting motivations.

Finally, I highlight two particular advantages of the general perspective promoted in this thesis. Firstly, the argument strategy pursued throughout the thesis reflects the attempt to understand the dynamics of those scientific activities which lead to the construction of explanatory models/theories of cognition. This in turn promotes the cultivation of a more acute sense of philosophical modesty that should guard against legislating in a purely *a priori* manner what should count as a proper explanatory account of a given class of empirical (cognitive) phenomena. Secondly, the concern for the normative dimensions of the proposed account of cognitive explanation corresponds to the complementary requirement that a philosophical analysis should provide a series of critical tools for analysing the underlying principles and assumptions of various scientific theorising and experimental practices. Together, these two ideas provide a sound basis for what I take to be an informative and adequate philosophical analysis of any problem which has also roots in the scientific domain.

2

LESSONS FROM PHILOSOPHICAL THEORIES OF EXPLANATION

2.1 INTRODUCTION

The main objective of this thesis is to provide a novel and adequate account of the notion of cognitive explanation which embodies both insights derived from the careful survey of the explanatory schemas utilised by practising scientists and the general desiderata articulated by traditional philosophical models of scientific explanation. The first perspective is intended to secure the descriptive adequacy of the account by reflecting both the achievements as well as the challenges that confront the scientific study of cognitive phenomena. The role of the second perspective, on the other hand, is to balance the fragmented picture of the explanatory frameworks used in the different disciplines of cognitive science by identifying their common features and principles. Thus, drawing on the lessons afforded by traditional analyses of the structure of scientific explanation, I articulate a view of cognitive explanation which mirrors the main aims of cognitive scientific research but also constitutes a critical tool for evaluating the proposed models/theories of particular cognitive phenomena. In addition, by combining these two perspectives, I elucidate the sense and extent to which cognitive explanations constitute a distinctive mode of scientific explanation. That is, the emerging account will also clarify what sets cognitive explanations apart from other types of scientific explanation.

The task of analysing the structure of cognitive explanation faces three important challenges. First, the analysis seems to be torn between two *prima facie* opposing requirements: (i) to acknowledge the variety of explanatory schemas utilised at different levels of analysis and abstraction by particular groups of cognitive scientists, and (ii) to identify any common features and principles underlying the practice of constructing explanatory theories/models of cognitive phenomena. The difficulty of this challenge resides in proving that the second requirement does not amount, in fact, to the reduction of a set of different explanatory schemas to a single general strategy which would be applicable in all the disciplines of cognitive science irrespective of the particular problems each of them addresses. As will be shown in the following chapters, this *explanatory monist* tendency is present in

all current philosophical treatments of the notion of cognitive explanation. In contrast, I attempt to show that the tension between these two requirements can be resolved without brushing aside the relevant features which distinguish the explanatory frameworks proposed in different branches of cognitive science.

Second, there is yet another type of diversity which seems to frustrate the prospects of a 'well-behaved' philosophical model of the notion of cognitive explanation. For, given the increased compartmentalisation and specialisation of the disciplines that study cognitive phenomena at various levels of analysis, abstraction, and resolution, the question arises whether there is any uniform way of delimiting the proprietary objects of cognitive scientific research from those of other sciences such as neurobiology, cellular and molecular biology, electrophysiology, and so on. One possible way of meeting this challenge is simply to admit the open-ended structure of cognitive scientific research which implies that its objects (the explananda of cognitive theories/models) are never fixed in an absolute way. Otherwise put, under this view, there is no uniform way of determining what counts as the explananda of a particular cognitive theory/model, independently of the actual scientific practice in which a particular problem is raised and investigated. Another possible strategy of circumscribing the proprietary explananda of cognitive theories, that has been more popular among philosophers of mind, consists in identifying a common feature (or a cluster of features) shared by all and only cognitive phenomena.[1] On this *conceptual* (or *a priori*) strategy, the objects of cognitive explanations are precisely the phenomena that can be said to have this distinctive feature. Although in what follows I will provide a number of compelling reasons for preferring the first strategy for dealing with this 'demarcation' issue, I also admit that it generates a much more heterogeneous basis for a potential general account of cognitive explanation.

The third challenge for a substantive account of cognitive explanation is to clarify what distinguishes *bona fide* cases of cognitive explanations from other types of scientific achievements and aims that are also legitimately pursued by practising cognitive scientists. That is, such an account must be able to justify the conceptual (logical) independence of the explanatory value/power of cognitive models/theories from their empirical adequacy, simplicity, and unificatory power, their ability to support novel predictions and/or yield testable experimental hypotheses. In addition, the account must motivate the idea that developing explanatory theories is a productive aim of cognitive science in the first place. I will show that these chal-

1 The category of *intentionality* has been traditionally taken to constitute such a distinctive feature of mental or cognitive phenomena (cf. Crane 2003). More specifically, Burge (2010) has argued that a philosophical analysis should identify the *constitutive conditions* for something being a particular type of cognitive capacity (e.g., perception, language comprehension, etc.).

lenges can be met by adopting a *practice-based* perspective on the problem of cognitive explanation. The adoption of this perspective entails the commitment to pursue an in-depth analysis of the explanatory frameworks used by particular research communities to elucidate certain aspects of cognitive phenomena at various levels of analysis or resolution.

Moreover, this approach does not start by assuming any particular definition of the cognitive domain, but rather lets itself be guided by the ways in which various investigative practices have been organised and developed in order to tackle different aspects of mental or cognitive phenomena. This avoids ruling out *a priori* (i.e., in advance of actual scientific inquiry) certain research programmes as being inadequate with respect to the purported proprietary objects of cognitive science. Instead, the practice-based viewpoint affords a more local analysis of the dynamics of cognitive science and of its varied explanatory aims, promising to yield a better understanding of the relationships holding, within cognitive science, between its component disciplines, as well as between these and other areas of scientific investigation. This is important because, as even a cursory look at the recent history of cognitive science shows, the remarkable development of the field was facilitated to a significant extent by the co-opting of various experimental and theoretical strategies from other areas of scientific research, as well as by the development of new techniques (e.g., PET, fMRI, and TMS scans). Understanding how these tools and techniques contribute to the construction of explanatory accounts of cognitive phenomena constitutes an important part of a robust philosophical account of cognitive explanation. In addition, by focusing on the practice of constructing and refining explanatory theories/models of cognitive phenomena, one is better placed to appreciate the often intricate relationship between the explanatory value of a cognitive theory and other properties such as its simplicity, unificatory power, etc.

Although the proposed strategy is not entirely novel, I claim that its application in this thesis will yield a distinctive view of cognitive explanation which differs in significant respects from other philosophical accounts. In developing this account, I argue that the philosophical problems that come up in connection with the modelling and theorising practices of cognitive scientists are not entirely specific to the cognitive domain and most of them are variants of problems that have been discussed in other contexts before. This is not to say that the disciplines of cognitive science do not raise new philosophical problems of their own. Rather, the point is to consider how the literature on cognitive science can contribute to existing debates about, among others, scientific modelling, idealisation, and explanation, rather than viewing it as exploring completely new territory. Furthermore, I claim that emphasising the continuity with existing debates in philosophy

of science increases the prospects of providing a better analysis of the problem of cognitive explanation itself.

In line with this strategy, section 2 represents a critical survey of three traditional accounts of scientific explanation. My aim is to show how these accounts have shaped current models of cognitive explanation and to discuss the general conditions/desiderata they impose on a philosophical account of explanation. Then, in section 3, I offer a brief characterisation of the factors that have generated some of the major research questions of cognitive science. I will show that there is a close connection between these factors and the currently prevalent philosophical models of cognitive explanation. Lastly, section 4 draws the general outline of the strategy that will be pursued in the rest of the thesis in order to articulate and refine a novel and adequate account of scientific explanation in the cognitive domain.

2.2 TRADITIONAL ACCOUNTS OF SCIENTIFIC EXPLANATION

I propose to develop the analysis of the structure of cognitive explanation against the background of traditional philosophical treatments of the notion of scientific explanation. There are three important motivations for starting the investigation from these general considerations. Firstly, this way of introducing the problem of cognitive explanation situates the main objective of the thesis more clearly in the area of philosophy of science and philosophy of explanation, thus differentiating it from other debates currently taking place within certain areas of philosophy of mind. Secondly, I take traditional philosophical models of scientific explanation to provide the most general set of conditions that will direct the construction of my account of explanation in cognitive science. And, thirdly, by analysing the limitations of the traditional accounts that have been most influential in shaping current philosophical models of cognitive explanation, I aim to identify a series of salient problems that an adequate account of cognitive explanation must address.

The philosophical accounts on which I focus in what follows are the *covering law* model of scientific explanation (Hempel and Oppenheim 1948; Hempel 1965, 2001), and two other views that have emerged as critical responses to the problems faced by the Hempelian model of scientific explanation: the *statistic/probabilistic* (Jeffrey 1969; Salmon 1971; Mellor 1976) and *causal* models of scientific explanation (Salmon 1984a, 1989; Lewis 1986; Psillos 2002; Woodward 2003). Whilst most conceptions of cognitive explanation have been defined in contrast to the covering law model of scientific explanation (e.g., Fodor 1974; Cummins 2000; Craver 2007b), the statistical/probabilistic and causal accounts have had a more positive influence, inspiring a range of views that seem to be adequate with respect to the cognitive domain. Thus, the task of the following three sections is to outline the main

tenets of these three traditional models of scientific explanation. In particular, I focus on the criteria that these accounts have put forward for qualifying something as a genuine scientific explanation. Also, I aim to identify the major challenges that these accounts still have to confront. In light of these critical analyses, I identify a number of important lessons that should guide the search for a substantive account of cognitive explanation.

Debates concerning the notion of scientific explanation have also given rise to several classificatory schemes that are supposed to distinguish the various accounts proposed in the philosophical literature. For instance, a number of authors have followed Salmon (1984a) in classifying accounts of scientific explanation into *metaphysical (ontic)*, *epistemic*, and *modal* views (e.g., Kitcher 1989; Craver 2007b; Wright 2012), whereas others have preferred Kim's distinction between *internalist* and *externalist* views of scientific explanation (e.g., Kim 1994; Batterman 2002). Another category that has been more recently promoted in the philosophical literature covers the various versions of *pragmatist* accounts of scientific explanation (e.g., van Fraassen 1980; Faye 2007; Ylikoski 2007). Rather than trying to fit any of the accounts analysed into the following sections in any of these categories, I prefer to discuss, when relevant, their distinctive metaphysical, epistemic, and pragmatist commitments. This strategy will avoid, I hope, confusing the criticism of a particular view with an objection raised against the label attached to it. As a final terminological point, I will use the designation 'epistemic' as relating to the human process of seeking knowledge, without implying that epistemic states are necessarily truth-bearing, and the term 'pragmatist' to designate whatever is related to particular types of (scientific) practices.

2.2.1 *Hempel's model of scientific explanation*

The primary aim of this section is to provide an analysis of the *covering law* model of scientific explanation (cf. Hempel and Oppenheim 1948; Hempel 1965) that will afford an instructive way of assessing the continuities between Hempel's classical approach and subsequent philosophical accounts of explanation that have shaped the landscape of models of cognitive explanation. A further objective of this analysis is to identify the most important insights that the Hempelian conception of scientific explanation makes available for the construction of an adequate account of cognitive explanation.

The core idea of the Hempelian approach is that explanations are essentially arguments or derivations. There are two components of the covering law model of explanation: *deductive-nomological* (D-N) explanations and *inductive-statistical* (I-S) explanations. According to Hempel (ibid.), most scientific explanations have the logical form of a deductive argument that shows that the explanandum (conclusion)

follows from a set of premises expressing general regularities or laws and certain initial and/or boundary conditions. The original source of the D-N model of explanation was provided by the case of solving initial value problems for linear differential equations. Since differential equations are involved in the investigation of a wide range of empirical problems involving the motion of fluids, the flow of current in electric circuits, the dissipation of heat in solid objects, the propagation of seismic waves, and the propagation of light, among many others, they were taken to constitute an appropriate base for a general model of scientific explanation. Thus, in the case in which laws are represented by ordinary differential equations, according to the Hempelian model, one has an explanation of some fact (for instance why the system is in state S_i at time t_i) if one can solve the equations given the system's state at some time t_i. More importantly, Hempel also implies that the derivation of the final state of the system (at some time t_j), from the laws and initial or boundary conditions should provide some insight or understanding of the phenomenon in virtue of the fact that the derivation enables one to see why the explanandum phenomenon was to be expected at all.

The generalisation of this particular type of context has been taken to support the idea that the explanation of a target phenomenon standardly consists in the *deductive* subsumption of the explanandum under the dynamical laws of the appropriate theory together with some contingent matters of fact, which constitute the initial and/or boundary conditions for the phenomenon under investigation. Moreover, Hempel (1965) has argued that the explananda of deductive arguments can be both occurrences of particular events (e.g., the appearance of a rainbow or of a lunar eclipse on a particular occasion) and general patterns expressed by natural laws themselves (e.g., the laws of refraction and reflection that are invoked in the explanation of rainbows can themselves be explained by the electromagnetic wave theory of light).[2] This in turn has been taken to imply the applicability of the D-N model to a wide range of scientific cases.

In the particular case of theories which appeal only to deterministic laws, the fact that the *explanans* sentences show that the *explanandum* phenomenon is to be nomically expected coincides with the idea that the explanation relation is actually one of logical entailment. In other words, in these cases, a particular theoretical description counts as an explanation of a particular phenomenon if it can be recast in the

2 Because of its insistence of the explanatory role of natural laws, the D-N model downplays the importance and sometimes the complications introduced by fixing certain (internal or external) boundary conditions in the explanation of phenomena such as the formation of rainbows. For an in-depth analysis of these examples and of the ways in which the fixation of 'initialandboundary' conditions in physical modelling can generate surprising shifts in explanatory structure, see Wilson (1992, 2010) and Batterman (2002, 2010).

form of a deductive argument where a specific set of premises (the *explanans*) logically entails the conclusion (the *explanandum*).[3]

However, the main tenets of the covering law model have been taken to apply equally well to the case of theoretical explanations of non-deterministic phenomena, in which case the covering laws are probabilistic or statistical laws. Hempel conceived statistical laws as having the form of conditional probability statements, viz. $Pr(G \mid F) = r$ (the probability of something being F given that G holds is r). Since probabilistic or statistical laws cannot be said to logically entail the phenomena or patterns to be explained, according to Hempel, non-deterministic phenomena are amenable only to Inductive Statistical (I-S) explanations. More specifically, Hempel argues that I-S arguments count as *bona fide* explanations to the extent and only to the extent that the value of r happens to be reasonably high (Hempel 1965, p. 390). In other words, the proposed explanans sentences have a genuine explanatory function if and only if they bestow *high probability* to the occurrence of the explanandum.

Both the D-N and the I-S versions of the covering law model of explanation have been criticised on a number of grounds. For instance, in the case of the statistical model, it has been pointed out that the proposed way of determining the explanatory value of a particular probabilistic argument is undermined by the fact that there is no non-arbitrary cut-off below which the argument fails to explain and above which it would count as genuinely explanatory. Otherwise put, given their purported inductive form, statistical explanations seem to be intrinsically 'ambiguous' insofar as it is *in principle* possible to find conflicting 'explanatory' I-S arguments: one showing the high probability of the occurrence of one event relative to one reference class, and another showing, relative to a more restrictive reference class, the low probability of that same event/phenomenon (*the epistemic relativisation problem* for I-S explanations). These and further issues with the original I-S model have led to the development of alternative treatments of the notion of statistical/probabilistic explanation, some of which will be considered in the following section (e.g., Jeffrey 1969; Salmon 1971; Mellor 1976; Railton 1978, 1981).

In addition to the specific problems threatening the I-S model of explanation, there are two more general puzzles that have called into question the adequacy of the Hempelian account of scientific explanation. On the one hand, the wide variety of counterexamples encountered in the philosophical literature strongly suggests that Hempel's

3 'Thus a D-N explanation answers the question '*Why* did the explanandum phenomenon occur?' by showing that the phenomenon resulted from certain particular circumstances, specified in $C_1, C_2, \ldots C_k$, in accordance with the laws $L_1, L_2, \ldots L_r$. By pointing this out, the argument shows that, given the particular circumstances and the laws in questions, the occurrence of the phenomenon *was to be expected*; and it is in this sense that the explanation enables us to *understand why* the phenomenon occurred' (Hempel 1965, p. 88).

analysis of scientific explanation does not actually provide necessary or sufficient conditions for what counts as a good scientific explanation. Most of these counterexamples fall under either the category of the *asymmetry of explanation puzzle* or the *irrelevant detail puzzle*, neither of which, it has been argued, admits of a satisfactory treatment on the classical Hempelian view (e.g., Salmon 1989; Kitcher 1989; Hansson 2006, 2007).

However, the most salient weakness of the covering law model of scientific explanation concerns the strong requirement that scientific explanations be construed as arguments which have an explicit logical structure. For it has been argued that most scientific explanations do not come in a ready-made deductive or even argumentative form. Although a committed deductivist might insist that all actual (as opposed to merely schematic) scientific explanations could be reconstructed so as to express deductive inferences, this solution would seem to solve only part of the problem. The focus on the relation of logical entailment still obscures the fact that scientific explanations are normally developed in a more progressive manner and that they often involve a series of mixed-level assumptions concerning the organisation and structure of the phenomena being targeted by a particular explanation. These general considerations seem to undermine Hempel's model of explanation which rests on the problematic assumption that all parts of the explanans are more or less in the same epistemic boat, in that once they are admitted for the purposes of one explanation, they cannot be discarded at a later point in the course of scientific investigation.

In consequence, the main lessons afforded by the covering law model of scientific explanation seem to be primarily negative. The model embodies the desiderata of having a set of general criteria that determine in all contexts and for all times what counts as explanatory. As such, the covering law model of scientific explanation can be taken to be committed to *explanatory monism*, the view that all scientific theories and/or hypotheses must conform to the same standard of explanatory 'goodness'. Despite the criticism of the Hempelian conception, the thesis of explanatory monism has been adopted by almost all successors of this account. Whereas Hempel (1965) identified the concept of natural law as the single explanatory structure appropriate for the study of physical phenomena, other theorists proposed different categories in order to ground the explanatory power/value of all scientific theories at all times.

Whilst I will criticise this underlying monist assumption in more detail in the case of philosophical models of cognitive explanation, it should be noted that this particular hypothesis seems to be responsible for most of the problems faced by the Hempelian conception of explanation. For I take it that insofar as one interprets the Hempelian account as the more restricted claim that natural laws sometimes play

the role of adequate explanatory structures, the account manages to capture an important component of the explanatory schemas used in certain areas of scientific investigation. Finally, the criticism of the Hempelian model allows one to appreciate better the difficulties of attempting to elucidate the nature and structure of scientific explanation by focusing exclusively on its logical form. In other words, the various counterexamples raised against the D-N and I-S models of scientific explanation strongly suggest that no logical structure will be able to fix once and for all what counts as explanatory in any domain of human (scientific) inquiry.

In what follows, I propose to survey briefly two other philosophical approaches that have grown out of the Hempelian project: (i) the *statistical* model of scientific explanation, and (ii) the *causal* account. Both accounts have proposed particular solutions to the main problems that threaten the plausibility of the covering law model. In doing so, they have generated distinctive conceptions of scientific explanation which continue to inform substantively current debates in the philosophy of explanation, underlying some of the most prominent philosophical models of cognitive explanation.

2.2.2 *Statistical explanations*

The common element of the two types of explanation subsumed under the covering law conception is that scientific explanations are viewed as arguments, deductive or inductive, showing that the phenomenon to be explained - the explanandum - was to be expected in virtue of the explanatory facts set forth in the explanans. Whilst deductive validity is an all-or-nothing affair, inductive support comes in degrees. That is, the premises of an inductive argument are said to lend more or less weight to the conclusion, so that one can speak of degrees of strength of an inductive inference. More precisely, according to Hempel's view, 'an inductive-statistical explanation has a degree of strength which designates the *inductive probability* conferred upon the explanandum by the explanans' (Salmon 1971, p. 8). Two more specific problems have led to the development of alternative accounts of the notion of statistical explanation: (i) the difficulty of characterising precisely the conception of inductive inference presupposed by Hempel's I-S model (Jeffrey 1969), and (ii) the idea that even low-probability phenomena are amenable to scientific explanation (Salmon 1971).[4]

Both Jeffrey (1969) and Salmon (1971) took seriously these problems and argued against the idea that statistical explanations should be conceived on the model of inductive arguments. They also de-

4 Since Hempel's I-S model requires that a statistical explanation must embody a high probability, it rules out the possibility of explaining phenomena which might occur despite the fact of being intrinsically improbable.

fended the intuition that uncertain phenomena are amenable to rigorous scientific explanation. Thus, rather than considering high inductive probability as a criterion for the explanatory value of statistical models/theories of empirical phenomena, Salmon (1971) proposed an alternative *statistical relevance* criterion. To say that a specific factor 'is statistically relevant to the occurrence of an event means, roughly, that *it makes a difference to the probability of that occurrence* - that is, the probability of the event is different in the presence of that factor than in its absence' (ibid., p. 11). Thus, in essence, the *statistical-relevance* model (S-R) claims that an explanation is 'an assembly of facts statistically relevant to the explanandum, regardless of the degree of probability that results' (ibid). On this view, Salmon claims, even highly improbable events admit of a perfectly legitimate explanation if the statistical relevance criterion is met.

Proponents of the statistical model of explanation offer further justifications for characterising statistical explanation along these lines. Since in this limited space I cannot do justice to the details of all these accounts, I propose a more indirect route of evaluating the S-R model of scientific explanation. I will focus on a version of the statistical model which rejects altogether the idea that scientific explanations have an essential inferential or argumentative structure. Railton's (1978, 1981) deductive-nomological-probabilistic (D-N-P) model of scientific explanation will therefore fall beyond the scope of the proposed analysis.

On Mellor's (1976) model of *probable explanation*: 'a (good) explanation raises or makes high its explanandum probability, p; and the more it does so (*ceteris paribus*) the better it is' (cf. Mellor 1976, p. 232). Mellor recognises that this proposal is ambiguous between saying that the job of an explanation is: (i) to raise p or (ii) to make p high. However, he also argues for the virtues of the second reading, on which the point of raising p (whether from an alternative or from a previous value) can be reduced to the requirement of making p high. Mellor (1976) calls this the main tenet (T) of the probabilistic/statistical view of explanation.

One *prima facie* advantage of thinking that T should constitute the core of an account of the notion of scientific explanation is that it accommodates the insights made available by the Hempelian model, whilst avoiding some of its most problematic consequences. Thus, Mellor claims that, despite the close and intuitive link between explanation and inference, one needs not equate the two in order to identify the proprietary function of an explanation. More strongly, he argues that the case for T's plausibility can be supported even without relying on the conceptual link between explanation and inference. For what is required by an explanation of a deterministic or indeterministic phenomenon is not more evidence, inference, or confirmation to tell us what happened: observation has already told us

that. Instead, we seem to want to know why something happened *at all*. As I will argue in the next section, one form of explanation that satisfies this criterion is causal explanation.

But sometimes suitable causal explanations are not to be had, either because an event may lack sufficient causes (e.g., radioactive decay), or they may not be discoverable (e.g., the onset of breast cancer), or their sufficiency might just not show up in the available setup (i.e., in the terms fixed by our best theory). In all these cases what happened might not have happened for all we can learn about the causal history of a certain event, however we cannot satisfy the request for an explanation just by citing causal information. That is, there are cases when causal explanations need to be supplanted by probabilistic explanations. Now, assuming with Mellor that the epistemic possibility of an explanandum's falsehood admits of degrees and that relative probability measures epistemic possibility, we can say that, *ceteris paribus*, (good) explanations raise the explanandum's probability relative to the complete explanans (or, which is the same, they reduce the epistemic possibility of the explanandum's falsehood). This in turn is taken to constitute a more refined form of tenet T stated above.

Mellor (1976) strengthens the probabilistic account of explanation by adding an extra constraint which connects the notion of explanation with that of truth. This constraint is intended primarily to avoid the trivialisation of probabilistic/statistical explanations. On any model of explanation that invokes the explanandum's probability, it seems that the explanans must do more than raise it. That is, the explanans must be true and must incorporate any suitable information that is statistically relevant in the sense that it affects the explanandum's relative probability. Since statistically relevant information can both increase and decrease the relative probability of the explanandum (cf. Salmon 1971), in order to secure T, Mellor (1976) argues that one must add to it a *truth-constraint* (S), which requires that the proposed true explanans relate only to a true explanandum. For an 'explanans that could relate to a false as to a true explanandum is no explanans at all' (ibid., p. 237). In short, an adequate model of explanation which lists all the features requisite for a successful explanans, including its relation to the explanandum, must not be indifferent to the truth-value of the explanandum.

In summary, the picture of statistical/probabilistic explanation entailed by the conjoined T&S thesis says that (good) explanations ought to raise the probability of the (true) *explanandum* relative to all the suitable relevant data which are available. However, under this view, if the information specified by the explanans fails to satisfy the high-probability (T) and/or truth (S) constraints, one is not forced to claim that what one has *must* be an explanation. Sometimes the request for explanation can be simply unwarranted or unsatisfied. In light of these observations, I take the probabilistic account summarised above

to state that sometimes scientific explanations of particular uncertain phenomena/patterns amount to showing that there is a significantly high statistical correlation between the proposed explanans and the given explanandum.

2.2.3 *Causal explanations*

Another important class of scientific explanations are distinguished by the fact that they show that there is a casual link between the explanans and the explanandum phenomenon. The task of this section is therefore to outline the causal model of explanation. For present purposes, I will mainly rely on the classical view of causal explanation (cf. Salmon 1984a; Salmon 1989; Railton 1978; Lewis 1986), which I distinguish from later mechanistic and counterfactual elaborations of the causal view (e.g., Machamer, Darden, and Craver 2000; Glennan 2002; Woodward 2003). The classical causal approach emerged from a series of attempts to secure the D-N model against a number of especially recalcitrant objections discussed in the literature on scientific explanation. In particular, the two key problems which seemed to admit of a straightforward solution on the causal approach were: (i) the asymmetry problem, and (ii) the problem of explanations that do not mention any specific natural laws.

Whilst the asymmetry of explanation puzzle suggested that the relation of logical entailment does not suffice to characterise the structure of (good) scientific explanations, the absence of law-like generalisations in a range of actual scientific explanations contributed to a weakening of the idea that all scientific explanations must fit the classical covering law model. According to a number of authors (e.g., Salmon 1971, 1984a; Railton 1978; Lewis 1986), the solution to both these problems consists in defining the explanatory relation that holds between particular scientific statements or models in terms of the notion of causation (Railton 1981; Salmon 1984a) or causal dependence (Lewis 1986).

On the causal account, the explanans displays explanatorily relevant information about the explanandum if it cites the 'right' causal information concerning the occurrence of the explanandum. In this way, the issue of explanatory relevance becomes one of determining which factors from the causal history (or nexus) of an event or phenomenon are relevant in a particular explanatory context. In other words, scientific descriptions of observable phenomena have an explanatory function only insofar as they describe the relevant aspects of the causal nexus in which the explanandum phenomenon is embedded. Although some supporters of the causal view of explanation (e.g., Salmon 1989; Woodward 2000) admit that there might be cases of non-causal scientific explanation (e.g., any physical explanation that appeals to conservation principles), stronger versions of the the-

sis argue that all scientific explanation is causal explanation. In what follows, I will focus on three major features of this general picture of causal explanation.

First, it has been claimed that the causal account of explanation provides a good compromise between descriptive and normative views of what counts as scientific explanation. On the descriptive side, the causal account is faithful to the predominance of causal talk encountered in daily scientific practice. On the normative side, causal accounts hold that any admissible scientific explanation must specify the relevant causal detail required, in any given context, to account for the occurrence of a specific event or phenomenon of interest. Defenders of this type of account claim that causally complete explanations are better viewed as an ideal, with *actual* scientific explanations citing only a fragment of the causal history responsible for the occurrence of the phenomenon under investigation (cf. Railton 1981). Hence, the causal account of explanation is taken to yield the following (weaker) normative criterion for explanatory power/value: good scientific explanations have to cite only a part of the relevant causal information concerning a target phenomenon. In short, defenders of the causal account invoke both the robust normative commitments of their view as well as general considerations about the predominance of causal talk in scientific descriptions to reinforce the idea that the causal account satisfies both normative and descriptive desiderata for a comprehensive philosophical analysis of scientific explanation.

Second, there is an important element of continuity between the causal account and the D-N model of explanation. The proponent of the causal account maintains that some piece of information counts as the explanans in relation to another piece of information – the explanandum – if and only if the two are connected via a genuine causal relation. However, this is compatible with the idea that natural laws capture/express the significant causal relations that in turn are taken to ground scientific explanation. On this scenario, D-N arguments would turn out to be just a perspicuous way of exhibiting the causal-explanatory relations that hold between the explanans and explanandum. Thus, whilst an advocate of the causal account need not be committed to the pervasiveness of D-N arguments, she might hold that deductive arguments may provide an appropriate way of presenting the relevant aspects of the causal network that contains the explanandum. Of course, what is essential on the causal account is not the D-N form of explanatory arguments, but rather the idea that causal relations generate explanatory knowledge and/or understanding.

Third, the causal approach implies that, irrespective of the ways in which one is able to identify the relevant causal information pertaining to a target phenomenon, the ensuing description will have an explanatory function only if it cites the right kind and amount

of causal relations. Thus, putting aside the pragmatic factors which determine what counts as the relevant causal information in a given context, the causal account claims that exhibiting the causal structure of the phenomena being investigated will be explanatory in a wide range of contexts. As such, the causal approach qualifies as a project that analyses the concept of scientific explanation in a way that is time-independent, and also independent of the branch of science in which particular causal explanations are being proposed.

There are a number of *prima facie* advantages to endorsing a causal account of explanation. Most prominently, the account promises to resolve both the asymmetry and irrelevant detail puzzles. With respect to the first puzzle, the causalist argument says that the relation between the explanans and the explanandum is asymmetrical because it reflects the intrinsic asymmetry of causal dependence relations. Since the explanans displays the causal factors or conditions that are responsible for the effect described in the explanandum, the relation between the two sides of the explanation relation cannot be reversed without contradicting the postulate that the cause (temporally) precedes the effect and not *vice versa*.

The solution to the irrelevant detail problem is less straightforward and seems to depend on the assumption that, in any given explanatory context, the investigators will be able to specify exactly the amount of causal detail required by the relevance requirement. This raises the question whether it is possible to circumscribe any uniform causal relevance criterion or whether the causal view is so permissive that it allows for multiple causal relevance criteria to ground genuine scientific explanations (cf. Kitcher 1989). Another possibility is to endorse a broadly metaphysical strategy and to claim that the causal information cited in a good scientific explanation is non-redundant because it corresponds to the *actual* or *real* causal structure of the world. However, taken at face value, both strategies seem to imply that there is a problem with determining a uniform and general causal relevance criterion. The reason I raise this point is that, whilst the causal account of explanation has often been portrayed as an ontic or metaphysical view of scientific explanation (Salmon 1984a; Craver 2007b, 2012; Strevens 2008), an adequate solution to the irrelevant detail puzzle seems to require a broadly epistemic strategy which acknowledges that what counts as causally relevant in any explanatory context is always relative to a set of epistemic interests and goals characteristic of a particular research programme.

Further refinements of the causal view of scientific explanation include two types of account which have been particularly influential as models of explanations developed in the special sciences such as the different branches of biology or even the social sciences. These are: (i) the mechanistic account (e.g., Machamer, Darden, and Craver 2000; Glennan 1996, 2002; Craver 2007b) and (ii) the counterfactual account

of causal explanation (e.g., Lewis 2001; Woodward 2000, 2003; Pearl 2000). In a nutshell, on the mechanistic approach, the causal relation which connects the explanandum and the explanans is analysed in terms of an appropriate underlying causal mechanism relating the two components of the explanation. Furthermore, on the standard systemic characterisation of mechanisms, these are viewed as complex systems, whose component parts and activities are organised in such a way that they exhibit the function performed by the system as a whole. Causal mechanistic explanations are a species of constitutive de-compositional explanation which reveals something important about the inner workings of particular observable phenomena targeted by the scientific investigation.

More precisely, mechanistic decompositions are taken to be explanatory when they reveal the *actual* mechanisms that underlie, maintain or support the phenomena being investigated (e.g., Craver 2007b). In light of these features, certain authors have pointed out that mechanistic explanations are essentially *local* or *particular*. That is, mechanistic explanations: 'show us how *particular* occurrences come about; they explain *particular* phenomena in terms of collections of *particular* causal processes and interactions - or, perhaps, in terms of noncausal mechanisms, if there are such things' (Salmon 1984a, p. 184, m.e.).

Another advantage of the causal/mechanistic account of explanation is that, unlike the Hempelian conception of explanation, it does not assume that the explanans must show that the explanandum was to be nomically expected. Rather, 'it shows what sorts of expectations would have been reasonable and under what circumstances it was to be expected. To explain an event is to show to what degree it was to be expected, and this degree may be translated into practical predictive behaviour.' (Salmon 1971, p. 79). Whilst defending the plausibility of low-probability explanations, Salmon also points out that drawing on the connection between explanation and prediction does not suffice to characterise the explanatory value of a scientific theory/model. For although an explanation of a particular event might indeed provide good and, perhaps, complete grounds for rational prediction concerning that event, the reverse does not seem to hold. That is, rational prediction is not a sufficient condition for claiming that one has an explanation.

Tying explanation to prediction-making does not completely solve the problem of characterising the relation of explanatory relevance between the information conveyed by the explanans and the explanandum. On the causal account, explanatory relevance is redefined in terms of the relation of causal relevance. However, as implied above, without a precise way of characterising the notion of causal relevance, the latter is still faced with a version of the *epistemic relativisation problem*. Nothing rules out the possibility that, on this account, something

that would count as causally relevant in a particular context, will turn out to be irrelevant in a different (perhaps more restrictive) context.

Moreover, if causal/mechanistic explanations are indeed essentially local or particular, then it seems that the view has an important blind spot in accounting for scientific explanations of general patterns or regularities. Since most scientific explanations are aimed at general patterns or repeatable phenomena, rather than at particular (or exceptional) occurrences, the local character of mechanistic explanations seems to put into question the wide applicability of the causal/mechanistic view itself.

Some of the difficulties faced by the causal/mechanistic account in characterising the causal relevance criterion have been addressed by counterfactual theories of causal explanation. According to the latter, causal relevance relations should be understood as a sub-class of counterfactual dependence relations (e.g., Lewis 2001; Woodward 2000, 2003). For instance, Woodward cashes out the idea of counterfactual dependence in terms of what he calls *what-if-things-had-been-different questions* or *w-questions* for short. That is, he claims that an explanation 'must enable us to see what sort of difference it would have made for the explanandum if the factors cited in the explanans had been different in various possible ways' (Woodward 2003, p. 11). Whilst Woodward construes his view of counterfactual dependence along manipulationist or interventionist lines, other authors have extended counterfactual accounts of explanation beyond the boundaries of causal explanation (e.g., Ylikoski 2007; Ylikoski and Kuorikoski 2010; Bokulich 2011), by avoiding construing counterfactual dependence in terms of the possible causal manipulations of the system.

Nevertheless, as an account of causal scientific explanation, the counterfactual model exhibits a number of advantages. Firstly, the counterfactual account receives some support from psychological theories which show that there is a strong connection between causal judgments/reasoning and the psychological processes involved in counterfactual reasoning tasks (e.g. Gopnik and Schulz 2007). Whether or not philosophical analyses are able to show that causal notions and judgments may be exhaustively analysed in counterfactual terms is still an open problem (cf. the classical discussion in Lewis 1986; see also Edgington 2011). There are, however, other grounds for believing that counterfactuals play an essential role in the explanatory strategies utilised in science. One straightforward reason has to do with the wide use of idealisation and abstraction in scientific explanation. Both are involved in experimental and theoretical contexts and rely heavily on counterfactual assumptions about systems that are organised and behave in slightly different ways to the actual ones. In light of these connections between scientific causal modelling and counterfactual reasoning, it seems that explanations of general pat-

terns or regularities are better handled on the counterfactual account of explanation than on the causal/mechanistic one.

Secondly, several authors have argued that counterfactual dependence relations are more pervasive than causal relations proper (e.g., Ylikoski 2007; Bokulich 2011). If this claim is correct, it opens the door for a more comprehensive account of explanation that is able to accommodate other forms of explanation that are *prima facie* non-causal. Other authors have suggested that counterfactual analyses help to elucidate some of the processes that are involved in causal reasoning (Psillos 2004), as well as why we often treat causal descriptions as explanatory in the first place (Woodward 2011). In brief, it seems that the counterfactual account makes available a number of resources that help clarify the epistemic dimension of the causal view of scientific explanation, providing additional insight into why we treat causal information as explanatorily relevant in the first place.

2.2.4 *Philosophical models of scientific explanation: insights and issues*

There are several important lessons that follow from the analysis of these three classical models of scientific explanation which will further guide the construction of an adequate philosophical account of cognitive explanation. First, the challenges confronting the *covering law* model of scientific explanation (Hempel 1965) highlight the limitations of focusing one's philosophical analysis solely on the *logical* structure of explanation. Furthermore, although the Hempelian approach seems to be correct in identifying the notion of *natural law* as a potential explanatory structure, the account turns out to be problematic due to its commitment to the thesis of explanatory monism which implies that only scientific theories that invoke natural (deterministic or statistical) laws qualify as having genuine explanatory power. The other two accounts analysed above attempt to overcome the main problems of the Hempelian conception and have arguably had a more substantive contribution in shaping the landscape of current models of cognitive explanation.

The *statistical/probabilistic* model of explanation (cf. Jeffrey 1969; Salmon 1971; Mellor 1976) draws attention to the fact that even highly improbable events are *in principle* amenable to scientific explanation. By criticising the exclusivist focus of the Hempelian account on the inferential structure of scientific explanations, the statistical model claims that sometimes explanations take the form of the statement of a relevant statistical correlation between a particular explanandum and the proposed explanans. As such, the statistical model attempts to accommodate in the landscape of potential explanatory structures a series of tools and techniques that have been developed in order to deal with especially 'recalcitrant' or unexpected physical phenomena/events. The statistical model of explanation seems to provide

an appropriate framework for analysing the explanatory value of dynamic systems approaches to cognition, Bayesian models and connectionist models of cognitive phenomena (e.g., Rumelhart, McClelland, and PDP Research Group 1986; Thelen and Smith 1994; Kelso 1995; Port and van Gelder 1995; Griffiths, Kemp, and Tenenbaum 2008; Guastello and Pincus 2009; McClelland et al. 2010).

The *causal* model of scientific explanation (e.g., Railton 1981; Salmon 1984a, 1989; Lewis 1986) provides a compelling case for the idea that causal knowledge often plays an explanatory role in scientific inquiry. However, I have argued that, if one is to properly justify the explanatory power of certain causal structures, one needs to develop a more sophisticated set of conceptual resources. Along these lines, the notions of *mechanism* and *counterfactual dependence* have been put forward as essential ingredients for articulating a more robust philosophical conception of causal explanation (cf., Machamer, Darden, and Craver 2000; Glennan 2002; Woodward 2003; Craver 2007b). In addition, I have pointed out that despite its strong intuitive appeal, the causal model of scientific explanation needs to confront the epistemic relativisation puzzle and offer a more detailed and precise picture of the factors involved in the construction and evaluation of specific causal models/theories that might be deemed to be genuinely explanatory.

This perspective in turn indicates that the standard ontic conception of causal explanation (e.g., Salmon 1984b; Strevens 2008; Craver 2012) needs to be supplemented by a more careful discussion of the epistemic and pragmatic dimensions of the causal explanations developed in different areas of science. By analysing several models of cognitive explanation that share the central assumptions of these two classical accounts of explanation, I aim to show that this latter shift of focus yields a more compelling solution to both the regress and explanatory relevance problems faced by any philosophical model of scientific explanation.

Finally, I would like to mention briefly two other approaches to the problem of scientific explanation. The *explanatory unification* account (cf. Friedman 1974; Kitcher 1981, 1989) equates the explanatory power of scientific theories with their unificatory power. Although the tendency to conflate the two epistemic virtues (i.e., explanatory and unificatory power) resurfaces in almost all philosophical analyses of the notion of cognitive explanation, I argue that there are good reasons for thinking that they are logically distinct epistemic virtues. The presence of one of them does not guarantee that the theory in question also possesses the other virtue. In fact, adopting the unification criterion as an index of explanatory power/value invites the objection that unification can be achieved at the price of superficiality or even triviality (e.g., van Fraassen 1980; Hansson 2006). Moreover, the main intuition associated with the explanatory unification account,

i.e., that subsuming many things under one general principle has an explanatory force, can arguably be interpreted as an alternative formulation of the covering law model of scientific explanation. Also, given the fact that sometimes compartmentalisation in science can be equally explanatorily efficient, *contra* explanatory unification, I maintain that unificatory power is better conceived as a separate epistemic virtue of certain scientific practices.

The pragmatist approach to the problem of scientific explanation (cf. van Fraassen 1980; Achinstein 1983; Bromberger 1992) highlights the fact that the construction and evaluation of explanatory scientific theories/models depends on the epistemic interests, goals, and aims of specific scientific communities or research programmes. More importantly, the approach draws attention to the fact that assessing the explanatory value of scientific models/theories is always relative to the particular problems or phenomena they were intended to account for. Although the pragmatist approach has often been taken to yield a sceptical stance towards the philosophical project of analysing the structure of scientific explanation, more recently the account has given rise to a number of constructive analyses which emphasise the double contrastive and erotetic structure of scientific explanation (e.g., Ylikoski 2007; Ylikoski and Kuorikoski 2010). However, it is also true that most of these accounts developed along broadly pragmatist lines borrow various conceptual resources from the other models of scientific explanation analysed in the previous section (in particular from the causal model of scientific explanation).

The investigations carried out in the following chapters embody the main insights of the pragmatist approach to the problem of scientific explanation. More specifically, by developing an analysis which pays special attention to the variety of factors that play a role in the construction of different types of scientific explanations, I follow the pragmatist's proviso of not conflating the notion of explanatory power/value with the other epistemic virtues of scientific models/theories: empirical adequacy, simplicity, unity, elegance, etc. (cf. van Fraassen 1980). In addition, the proposed approach circumvents the premature commitment to *explanatory monism* implicit in almost all mainstream philosophical accounts of the notion of scientific explanation. Whilst explanatory monism is usually taken to support the project of constructing a general philosophical model of scientific explanation, I will develop an alternative approach that is equally compatible with the aim of elucidating the nature and structure of scientific explanation.

2.3 INTRODUCING THE PROBLEM OF COGNITIVE EXPLANATION

The main objective of this thesis is to advance a philosophical account of cognitive explanation that is both novel and adequate with respect

to the current explanatory frameworks used in the different branches of cognitive science. The strategy adopted for this purpose consists in developing a critical analysis of some of the most prominent models of cognitive explanation discussed in the philosophical and cognitive scientific literature in order to reveal both their underlying theoretical assumptions and their relationship with the actual scientific activities which generate potential explanatory theories/models of particular cognitive phenomena. The broader framework for investigating the structure of cognitive explanation is provided by the three philosophical models of scientific explanation analysed in section 2. Thus, the plan is to draw on the lessons afforded by these critical analyses in order to articulate a philosophical account which vindicates the intuition that explanation is an important epistemic goal of cognitive scientific research.

The proposed approach, therefore, combines two complementary perspectives that contribute in distinctive ways to the construction of a general/substantive account of cognitive explanation. The perspective that has already been introduced in the previous section serves two main purposes: (i) to lay down the principal desiderata for a philosophical analysis of the notion of scientific explanation and (ii) to highlight/emphasise the continuity between the problem of scientific explanation and the more specific problem of cognitive explanation. I have also claimed that the second implication is beneficial because it mitigates the claim that the domain of cognitive science raises a number of completely novel problems within philosophy of science, by showing how various ideas and insights from the latter domain can be brought to bear on the analysis of the specific issues raised in the philosophy of cognitive science. In addition, this continuity claim opens up the reverse possibility: that hypotheses and insights made available by analysing various problems which arise in the domain of cognitive science might apply to other areas of scientific inquiry. By focusing on the development of a substantive account of cognitive explanation, this thesis raises the question of what constitutes an appropriate philosophical approach to the notion of scientific explanation.

The preceding analysis of the three classical models of scientific explanation has also revealed the limits of relying solely on this perspective for the purposes of developing an adequate account of cognitive explanation. Because all the accounts analysed in the previous section are committed to the thesis of explanatory monism, they run the risk of reducing prematurely the diversity of explanatory frameworks that are required in order to elucidate various cognitive problems. This limitation becomes particularly salient when one adopts the complementary practice-based perspective which mirrors the multiplicity of explanatory schemas used in the different disciplines of cognitive science. The fact that explanatory monism has been widely taken to em-

body the normative dimension of scientific explanation also explains why it tends to resurface in most, if not all, accounts of cognitive explanation. In contrast, I claim that explanatory monism is not the only way to vindicate the intuition that explanation (in any domain of inquiry) has a normative bite.

A practice-based perspective is not incompatible with the requirement that there must be a way to distinguish between explanatory and non-explanatory theories/models of cognitive phenomena. However, by adopting this perspective one becomes more aware of the fact that the construction of particular explanatory models/theories is dependent upon a wide range of factors, not all of which are under the researcher's willful control. Sometimes the complexity of the phenomena being investigated, together with the limitations (practical, technical, and intellectual) of the practicing scientists require the development of very ingenious strategies for advancing one's understanding of the phenomena being investigated. And, of course, sometimes these efforts might remain completely unsatisfied. Thus, the practice-based perspective is also intended to vindicate the intuition that part of the difficulty of saying anything precise about the notion of cognitive explanation derives from the difficulty of circumscribing the phenomena being investigated in the different fields of cognitive science.

Whilst the practice-based perspective I propose to adopt throughout the thesis seems to place one in the middle of things, I argue that it does not preclude the formulation of a general philosophical analysis of the notion of cognitive explanation. The purpose of the following considerations is to show that, although a practice-based perspective has been implicit from the very beginning in debates concerning the status and structure of cognitive explanation, the explanatory monism thesis imported from classical accounts of scientific explanation, and, more generally, the prioritisation of the problem of the normative character of explanation, have tended to obscure some of the most important insights that the adoption of this perspective makes available.

2.3.1 *A bit of history*

This section proposes a short detour through the recent history of cognitive science in order to show how the plurality of explanatory frameworks currently used in its different sub-branches is related to the main factors that have contributed to the constitution of the field. First, the field of cognitive science emerged as a strong reaction to *behaviourism*, the dominant research strategy ruling psychological studies at the beginning of the '50s.[5] The basic assumption of

5 More precisely, there are at least three major views which have been associated in the psychological and philosophical literature with *behaviourism*: radical, analytical

the behaviourist programme was that all mental phenomena or patterns should be explained without any appeal to unobservable mental states, relying instead on non-psychological mechanisms linking particular stimuli (inputs) with particular responses (outputs). Furthermore, these mechanisms were taken to be the product of conditioning. Among the various problems encountered by behaviourism as a general research strategy in scientific psychology, were two key issues that progressively led to the demise of the radical behaviourist programme. Firstly, behaviourism tended to isolate scientific psychology from other disciplines by denying the pertinence of the hypotheses and tools developed in other scientific domains for the study of the mind (or psychological behaviour). Secondly, an increasing number of acute theoretical analyses and ingenious experiments revealed an impressive range of cognitive or psychological behaviours that cannot be adequately explained in terms of stimulus-response mechanisms (e.g., Tolman, Ritchie, and Kalish 1946; Broadbent 1954; Miller 1956; Chomsky 1957, 1959).

There are two important lessons to be drawn from thinking of cognitive science as a critical response to behaviourism.[6] On the one hand, behaviourism showed the limits of adopting a single narrow methodology for the study of a very wide and diverse range of psychological phenomena. Thus, the criticism of the monist methodological commitments of behaviourism opened the possibility that different psychological phenomena or patterns might be appropriately investigated and explained with the help of different sorts of experimental and conceptual tools. Furthermore, the criticism of the radical behaviourist programme revealed some of the mistaken assumptions behind the exaggerated reaction against postulating unobservable abstract mental structures for the explanation of mental or cognitive phenomena. There is an additional cautionary lesson that follows from the fact that psychology's move from behaviourism was a lengthy and drawn-out process (which, according to some, has

(or logical), and methodological behaviourism. The distinctions between these positions can be characterised in the following way. Radical behaviourism corresponds to the project of explaining all mental phenomena in terms of stimuli, response, and reinforcements, without any appeal to mentalistic vocabulary. Among some of its most famous promoters, one can count: Edward Thorndike (1875-1949), John Watson (1878-1958), Ivan Pavlov (1849 - 1936), and B. F. Skinner (1904-1990) and Clark Hull (1884 - 1952). Logical behaviourism is a 'semantic' project which seeks to analyse all mental vocabulary in terms of stimuli and (dispositions to) response (Ryle 1949/2002). Finally, methodological behaviourism expresses the commitment that scientific psychology should concern itself only with the external behaviour of organisms (see also Rey 1997).

6 A pertinent historical observation in this context is that George Miller and Jerome Bruner (founders of the Harvard Center for Cognitive Science) originally introduced the term 'cognitive science' to designate the disciplines that studied not just the smaller subset of rational processes, but all mental phenomena. The choice of the word 'cognition' was merely supposed to distinguish their new approach from 'behavioural (non-mental) psychology'.

not yet been (or cannot be) completed). For despite its many faults, behaviourism rests on a strong and compelling intuition, namely that experience plays an important role in shaping various domains of cognitive processing at different levels of organisation. As will be shown in the following chapters of the thesis, this intuition continues to motivate a number of current research programmes in cognitive science which further testifies to the persistence and force of certain behaviourist assumptions in the study of mental phenomena.

A second striking feature that has characterised the field of cognitive science from its very beginning is its interdisciplinary nature. A rough picture of some the most influential ideas that triggered the wide range of research programmes currently pursued in cognitive science covers fields as diverse as mathematical logic, linguistics, neurobiology, cybernetics, and different branches of traditional scientific psychology (developmental, social, and evolutionary psychology, etc.). Among some of the most influential notions that have shaped the field of cognitive science, one could count: (i) the idea of algorithmic computation in mathematical logic (Turing 1937), (ii) the emergence of linguistics as the formal analysis of language (Chomsky 1957), and (iii) the progressive introduction of information-processing models of specific psychological processes in various sub-domains of cognitive science (Broadbent 1954; Miller 1956). Other important influences include the contributions of the different sub-fields of neurobiology which more recently have come to dominate the space of cognitive modelling and theorising.[7]

The last general feature that I would like to sketch briefly here concerns the predominance of information-processing models in cognitive studies. Even with the recent ascendancy of neurobiological models, one of the most prominent notions that continues to be invoked in a host of explanatory contexts throughout cognitive science is that of *mental representation*. Broadly speaking, there are two main sources of this widespread representational talk in cognitive studies. On the one hand, mental representations are central to pre-scientific folk psychology, even though folk psychology fails to provide a rigorous definition of the notion of representation. On the other hand, mental representations came to be understood along the lines promoted in computer science, as symbols in an information processing system (such as a digital computer). This latter characterisation of mental representations highlights two properties of these theoretical entities postulated in the explanation of cognitive processes/phenomena, namely that: (i) they refer to things outside the system, and (ii) they enter into symbol processing operations.

7 Although until the 1960s many cognitive scientists believed that the mind could be studied without studying the brain, the development of new technologies (such as PET and fMRI scans) for studying neural activity and new ways of modelling neural systems (e.g., artificial neural networks and cellular automata) reinforced the idea that the study of the mind is intimately related to the study of the biological brain.

These two ideas have played a major role in much philosophising about information processing or computationalist approaches to cognition. As a consequence of this widespread reliance on the notion of mental representation in cognitive studies, much of the effort in recent philosophy of mind has been aimed at clarifying the assumptions and implications of this broadly representationalist picture of the mind. Each of the following chapters of the thesis attempts to shed further light on the role(s) that mental representations play in the construction of explanatory accounts of various types of cognitive phenomena. However, the task thus circumscribed should be clearly distinguished from another major philosophical project which seeks to construct a substantive theory of mental content (e.g., Field 1978; Millikan 1984; Block 1986; Dretske 1988; Fodor 1987; Papineau 1987 etc.). In other words, this thesis seeks to develop an analysis of the role(s) played by the notion of mental representation in cognitive theorising and experimentation whilst remaining neutral with respect to the prospects of theories of content.

2.3.2 *Explanatory paradigms in cognitive science*

The previous considerations regarding the emergence of the domain of cognitive science reflect the fact that the class of explanatory frameworks used to investigate cognitive phenomena was and continues to be very heterogenous. The multiplicity of explanatory schemas used by practicing cognitive scientists mirrors both the interdisciplinarity of the domain, and the variety of cognitive phenomena currently investigated in different branches of cognitive science, at different levels of analysis or abstraction. Nevertheless, this fragmentation of cognitive studies raises two further interrelated questions: (i) are the explanatory schemas used to gain understanding of interesting cognitive phenomena the same as the ones used in other areas of scientific investigation?, and (ii) are cognitive explanations distinguished from other types of scientific explanations in virtue of the objects which constitute their explananda?

In response to the first question, a number of authors have argued that traditional philosophical models of scientific explanation, such as the covering law model, are inappropriate to describe and elucidate the explanatory practices encountered in cognitive science (e.g., Fodor 1968, 1974; Cummins 1983, 2000; Pylyshyn 1984). For instance, Cummins (2000) has pointed out that one of the main reasons for doubting the applicability of classical models of explanation to the cognitive domain is that the latter counts as explananda a different class of things than other natural sciences.[8] More specifically, he argues that because cognitive science deals with a distinctive restricted

8 The standard contrast class is that of theoretical explanation in physics (Fodor 1974; Cummins 1983). Despite the differences that separate the two styles of scientific

region of the empirical world, the type of laws which characterise its proprietary domain are merely laws *in situ*. They specify effects or regular behavioural patterns which are characteristic only of a specific type of system (i.e., cognitive systems). As such, psychological or cognitive laws are to be contrasted with the kinds of laws postulated by physical theories which hold for a wider variety of physical systems, under a large range of varying conditions.

The assumption underlying the contrast between cognitive and physical explanation seems to be that physical laws can play a genuine explanatory role in virtue of their generality. That is, since they are not (standardly) used to characterise specific types of systems, the primary role of physical laws cannot be that of identifying distinctive effects of those systems. Instead, within cognitive science, Cummins claims that '[one] should seek to discover and specify the effects characteristic of the systems that constitute their proprietary domains, and to explain those effects in terms of the *structure* of those systems, that is, in terms of their constituents (either physical or functional) and their modes of organization' (Cummins 2010: 288).[9] In addition, Cummins (2010) points out that effects (or laws *in situ*) are not the sole explananda of cognitive science. In fact, he claims that the primary explananda of cognitive research are psychological or cognitive capacities (e.g., the capacity to see depth, to learn and speak a language, to predict and make decisions, etc.). However, since a capacity is a kind of complex dispositional property, it also follows that to have a dispositional property is to satisfy a law *in situ*, i.e., 'a law characterising a certain kind of thing' (ibid). On this sort of account, the main difference between a cognitive capacity and an effect (or a law *in situ*) is that the former is usually harder to specify than the latter. Moreover, whereas cognitive effects usually need to be discovered (often through sophisticated experimentation), cognitive capacities have an intuitive pre-theoretical characterisation.

Despite these differences, Cummins (2010) argues that there is a general explanatory framework that is appropriate with respect to both types of explananda of cognitive theories/models (i.e., effects and capacities). This explanatory framework is known in the literature as *interpretative* or *functional analysis* (e.g., Cummins 1983). Functional analysis consists in decomposing a given cognitive capacity, identified as problematic, into a number of less problematic subcapacities such that the organised manifestation of these *analysing* sub-capacities amounts to a manifestation of the target analysed capacity. Such an analysis is said to *explain* how a particular complex system as a whole (i.e., a cognitive system) exercises the *analysed* capacity by showing it to be the result of the organised exercises of the

explanation, the polarisation promoted from within philosophy of mind and/or philosophy of psychology can be, or so I shall argue, misleading (cf. Wilson 1985).

9 See also Cummins 1983: chapters 1 and 2 for a more extensive discussion of how this kind of explanation is supposed to apply to psychology.

simpler *analysing* sub-capacities. In the case of certain systems, functional analysis goes hand in hand with the *componential* analysis of the target system. In such cases, the *analysing* sub-capacities are the capacities exhibited by the components of the system under investigation. However, Cummins (1983, 2010) and other authors have pointed out that this form-function correlation is often absent when analysing complex cognitive systems. This in turn has been taken to entail the relative autonomy of functional analyses of cognitive capacities from the componential analyses of the underlying biological mechanisms that support them (cf. Fodor 1974; Cummins 1983).

Although functional analysis has a large number of applications within cognitive science, as an explanatory framework it also has a series of limitations. For instance, the explanatory role of functional analysis seems to be constrained by the following factors: (i) the extent to which the *analysing* capacities are less problematic than the *analysed* capacities, (ii) the extent to which the *analysing* capacities are different in kind from the *analysed* capacities, and (iii) the relative complexity of the organisation of the component parts or processes that is attributed to the system (e.g., Cummins 2000; Egan and Matthews 2006). A further concern that challenges the adequacy and/or sufficiency of functional analysis as an explanatory framework for cognitive science derives from the idea that a complete theory of a cognitive capacity must also exhibit details of the target capacity's realisation in the biological system (or system type) that has it. That is, the implicit assumption is that 'the functional analysis of a capacity must eventually terminate in dispositions whose realisations are explicable via analysis of the target system. Failing this, we have no reason to suppose we have analysed the capacity as it is realised in the system' (cf. Cummins 2010: 292).

The primary motivation for starting the investigation of the explanatory frameworks used in cognitive science with the case of functional analysis is that it can be shown to underlie several prominent models of cognitive explanation. This further implies that some of the major problems facing the explanatory strategy associated with functional analysis will carry over to these other explanatory frameworks as well. Among these one may include: (i) belief-desire-intentions explanations (widely used in developmental and social psychology), (ii) classical or symbolic computational explanations, (iii) connectionist explanations, (iv) evolutionary explanations, and (v) neuroscientific-based or mechanistic explanations (cf. ibid.). Since the following chapters of the thesis will pursue an in-depth analysis of some of these influential explanatory frameworks, I propose to restrict my discussion here to outlining two foundational problems that arise for functional analysis generally, and that resurface in the context of other influential explanatory frameworks as well.

2.3.3 *Two challenges for cognitive explanation*

The two general problems that seem to affect all explanatory frameworks currently used within cognitive science are: (i) the *realisation problem*, and (ii) the *unification problem*. The first problem arises because the explanatory frameworks which share the characteristics of functional analysis seem to leave a gap between the functional characterisation of a cognitive system and the various nonfunctional descriptions that are taken to apply to the same cognitive system at a different level of analysis (e.g., neurobiological characterisations). The second problem amounts to the challenge of offering a unified account of cognition. That is, an account that postulates a set of general principles which underlie all types of cognitive processes, from early vision and motor control to higher-order processes such as language production and comprehension, reasoning, and decision making.

The realisation problem has often been invoked for the purposes of either highlighting or undermining the advantages of certain explanatory frameworks over their alternatives. In opposition to this line of reasoning, I seek to show that most versions of the problem are actually orthogonal to the task of characterising the structure of cognitive explanation. The main reason why the realisation problem has been taken, at least in philosophical circles, to constitute such an important cornerstone for explanatory accounts of cognitive capacities is the excessive concern with the nature of evidence and reliability in connection to the problem of explanation. However, as pointed out in the first part of this chapter, there are good grounds to resist the conflation of these different notions. By analysing in the thesis several versions of the realisation problem, I aim to show that there is no satisfactory formulation of the problem that is amenable to a general and informative treatment that would also impact the account of explanation. Thus, whilst the realisation problem might be interesting from a more metaphysical point of view, I claim that, in some of its most general forms, it tends to obstruct the philosophical analysis of the notion of cognitive explanation.

The unification problem, on the other hand, challenges the lack of unification of the explanatory frameworks used within the different fields of cognitive science. Those who are not content with the lack of unification of cognitive scientific theories usually want to allow that there might be different ways to investigate different aspects of the same cognitive capacity. However, they also point out that the deeper problem of cognitive science is that it fosters multiple incompatible explanatory accounts of the *same* cognitive capacity (e.g., language comprehension and production). The strong polarisation of some of the explanatory frameworks currently used within cognitive science (e.g., symbolic computationalism and connectionism, functional analysis and mechanistic decompositions) has been taken to reflect this

deeper disunity problem (cf. Cummins 2010). Moreover, in addition to the disunity encountered across frameworks there is considerable disunity within each framework. That is, one often finds competing models of the same cognitive capacity within the same explanatory framework. This situation seems particularly puzzling especially if one is inclined to think that cognitive capacities constitute distinct natural kinds, corresponding to determinate ontological categories.

In the following chapters, I attempt to dispel part of the worry generated by the unification problem by developing a piecemeal approach to the problem of cognitive explanation. The first step of this strategy is to provide some compelling motivations against taking unification to be the primary goal of cognitive modelling and theorising. I also aim to show that the notion of unification can be used to shed light on the intricate relationships between the different explanatory frameworks used within cognitive science. For this purpose, I develop, defend, and refine an account along the lines of the integrative pluralist position promoted by several authors in other areas of philosophy of science (e.g., Mitchell 2003, 2012; Chang 2012). One prominent advantage of this sort of position is that it promises to accommodate both competing and compatible alternative explanatory accounts of cognitive capacities. At a more general level, by showing how the pluralism of explanatory frameworks can be compatible with a notion of *local* unification, I seek to reinforce the claim that explanatory power and unificatory power are conceptually distinct epistemic properties of scientific theories.

2.4 OUTLINE OF THE STRATEGY

In light of the considerations put forward in this chapter, I maintain that a philosophical model of cognitive explanation should take into account two distinct factors: (i) the main lessons afforded by the critical analysis of traditional models of scientific explanation, and (ii) the multiplicity of explanatory frameworks currently being used within the different branches of cognitive science. Taken at face value, adopting a strategy that incorporates both factors might seem problematic because the two components pull in opposite directions. On the one hand, as we have seen, most traditional accounts of explanation seem to be committed to the thesis of explanatory monism, according to which there is a single normative core common to all the modes of explanation utilised across different scientific practices. On the other hand, even a cursory glance at the explanatory frameworks used by practicing cognitive scientists seems to suggest that a practice-based perspective will be able to license only a very permissive form of explanatory pluralism.

In consequence, the main challenge that confronts the present project arises from imposing two apparently incompatible requirements or

conditions on a philosophical account of cognitive explanation. These are the *descriptive adequacy* and *normative completeness* conditions. For it seems that an account which respects the descriptive adequacy condition is bound to accept the ineliminable diversity of the explanatory frameworks used to elucidate different cognitive phenomena. And this in turn is likely to frustrate the search for a uniform and prescriptive account of what counts as a reliable scientific explanation in the cognitive domain. In fact, the only type of normative principles which seem to be compatible with the descriptive adequacy requirement are the pragmatic principles which are inextricably connected to the aims, purposes, and interests of the agents involved in the relevant explanatory practices. Whilst I do not wish to dismiss the importance of the latter type of norms, I argue that they are not *in principle* inconsistent with a general picture of what makes scientific explanation a distinctive type of epistemic achievement. Thus, the task of the following chapters is to show that the two requirements are not actually incompatible, but rather that they have distinctive contributions to make to the formulation of a substantive philosophical analysis of the notion of cognitive explanation.

The following chapters seek to elucidate the main features of several explanatory frameworks used in the domain of cognitive science. The plan is to pursue an in-depth analysis of certain models of cognitive explanation that have been extensively discussed in the methodological philosophical literature and that have been taken to capture the dynamics of the explanatory strategies currently used in cognitive science. I begin with the mechanistic view of cognitive explanation because it constitutes an interesting case study in which classical philosophical analyses of scientific explanation meet the explanatory practices of certain communities of cognitive scientists. Next, I consider three different conceptions of computationalist explanations of cognitive phenomena whose main tenets are standardly taken to be in tension with one another. The arguments I put forward in chapters 4, 5, and 6 attempt to mitigate the strong polarisation of these accounts.

Although computational approaches to cognition do not, by far, exhaust the range of frameworks and/or strategies developed to investigate cognitive phenomena, I claim that their critical analysis will yield valuable insights on which to base an appropriate philosophical account of the notion of cognitive explanation that would apply beyond the boundaries of computational theories of cognition. The focus on mechanistic and computationalist theories/models of cognitive phenomena also reflects the idiosyncrasies of the more recent philosophical tradition which has been fascinated with the applicability of these general concepts to the study of the mind. I think that this particular fascination is in part explainable by the controversial metaphysical views that these concepts have encouraged. However,

since this investigation is not directly animated by the solution of any particular metaphysical puzzle concerning the nature of the mind, I propose to justify the focus on mechanistic and computational models/theories of cognition and their associated explanatory strategies in a way that is closer to the central theme of the thesis.

One general issue which arises in connection with the problem of cognitive explanation is whether abstract (e.g., mathematical, computational) models/theories can be taken to provide *bona fide* explanations of particular aspects of cognitive phenomena or whether concrete (e.g., mechanistic) models/theories are the only ones that are fit to fill this explanatory role in the cognitive domain. My proposal is that by paying closer attention to the details of the applicability of both concrete and abstract models and/or principles to the study of cognitive phenomena, one is in a better position to appreciate both the difficulties and the successes of using a wide variety of explanatory schema in order to have a better intellectual grasp of particular cognitive phenomena. Thus, the adoption of the practice-based perspective promoted throughout the thesis is not incompatible with the prospect of deriving a more general picture of the common features that characterise the practice of developing good explanatory accounts of cognitive phenomena. As a critical tool, attention to actual scientific practice checks certain sudden leaps and impulsive (or too optimistic) conclusions that philosophers tend to make concerning the structure of scientific scientific explanation in a domain such as cognitive science.

3

THE MECHANISTIC ACCOUNT OF EXPLANATION

3.1 INTRODUCTION

Current research in the fields of cognitive science and neuroscience makes use of a wide variety of models and techniques (e.g., Gazzaniga 2000; Shadmehr and Wise 2005; Stainton 2006; Sun 2008). Some of the most successful and influential of these rely on quite distinct theoretical assumptions about the structure of cognitive systems, as in the case of symbolic computational models, neural connectionist networks, and Bayesian models. The multiplicity of methods and tools used to investigate cognitive phenomena gives rise to the problem of how to evaluate the relative merits and limitations of the emerging scientific models/theories. As pointed out in the first chapter, one of the most prominent criteria put forward for dealing with this question consists in comparing the *explanatory power* of the various competing models and frameworks.

In this and the following three chapters, I will investigate a series of philosophical models of the notion of cognitive explanation that promise to shed further light on the varied landscape of theoretical and experimental approaches currently utilised in the area of cognitive science. The aim of these critical analyses is twofold. Firstly, I seek to determine the type of explanatory strategy characterised by each of these philosophical accounts and show how each is rooted in the scientific activities pursued by various groups of cognitive scientists and/or neuroscientists. Thus, each chapter begins by surveying the principal insights made available by a particular prominent model of cognitive explanation. Secondly, I identify the main challenges facing such philosophical models that attempt to characterise (often) in a uniform and general way the structure of various explanatory models/theories of cognition. This latter part of the investigation identifies the limitations of the available models of the notion of cognitive explanation and points towards an alternative, more fruitful, way of approaching the question of the explanatory value of specific cognitive models/theories.

In chapter 2, we have seen that the 'received view' of *cognitive explanation* tends to be constructed in opposition to classical philosophical accounts of the notion of *scientific explanation*. The driving intuition

behind this conception is that the cognitive domain raises a number of special issues that cannot be adequately tackled in the more traditional frameworks developed within philosophy of science (Cummins 1983, 2000). More recently though, *mechanism* has been put forward as an adequate account of cognitive explanation which manages, nevertheless, to preserve the most important features of traditional philosophical treatments of the problem of scientific explanation. Precisely because it promises to provide this important connection between classical and more specialised models of scientific explanation, I shall begin my investigation with *the mechanistic model of cognitive explanation*.

The *New Mechanists* (e.g., Glennan 1996, 2002; Machamer, Darden, and Craver 2000; Woodward 2000, 2003; Craver 2007b; Bechtel and Richardson 1993/2010) promise to deliver a robust notion of scientific explanation that has two highly desirable features, which have also partly motivated the extension of mechanism to the domain of cognitive science. Firstly, mechanism has been said to represent closely the explanatory and experimental practices encountered in certain scientific domains, such as the various sub-branches of cognitive psychology and neuroscience. Secondly, the account proposes a uniform and general set of criteria for something to count as having *genuine* explanatory power. As such, the mechanistic view arguably allows for the comparison of potentially competing explanatory models of a target cognitive phenomenon.

As a general account of explanation in cognitive science, mechanism holds that a *bona fide* explanation should exhibit the causal mechanisms that underlie, maintain, or produce the phenomena under investigation. If it fails to do so, a particular model or hypothesis cannot be said to have a *genuine* explanatory function. As a consequence, some defenders of the new mechanistic view subscribe to the idea that abstract (i.e., mathematical) models of cognitive capacities lack genuine explanatory power altogether (e.g., Kaplan and Craver 2011; Kaplan 2011). In other words, since explanation arises only if one can display certain aspects of the causal mechanisms underlying a particular phenomenon, mathematical (acausal) models are deemed to be inappropriate candidates for explanations.

The arguments developed in this chapter pursue two interrelated aims: (i) to offer a critical assessment of the main tenets of the new mechanistic philosophy of explanation, and (ii) to evaluate the advantages and limitations of applying the mechanistic framework to different sub-branches of cognitive science. I claim that, despite its intuitive appeal and its strong insights about some of the explanatory tools used within cognitive science, the mechanistic view faces a series of problems that call into question its supposed unrestricted scope and applicability. These challenges concern both the internal consistency of the mechanistic criterion for explanatory value, as well

as the emerging mechanistic position towards the status of abstract models.

The argument strategy I pursue is structured in three distinct parts. I begin, in section 2, by analysing the main tenets of the mechanistic picture of explanation. In section 3, I identify some of the most significant challenges facing the conception of mechanistic explanation and analyse three distinct strategies that mechanists might deploy to avoid these problems, finding the solutions they offer to be unconvincing. I support these objections to the general mechanistic strategy with respect to the cognitive domain by appeal to a series of models whose uses and functions in the context of cognitive neuroscientific research are analysed in detail in section 4.

The last section draws the main lessons from the critical analysis of the mechanistic view of cognitive explanation. In particular, I take the analysis to show that, contrary to the resolute version of the mechanistic thesis (e.g., Craver 2007b; Kaplan and Craver 2011; Kaplan 2011), abstract models are fit to fulfil proper explanatory roles even when they do not specify details of the causal/mechanistic structure in which a particular cognitive phenomenon is embedded. I conclude by outlining a restricted notion of mechanistic explanation that is more adequate with respect to the varied landscape of explanatory strategies encountered in cognitive science. The proposed interpretation vindicates the numerous insights that the mechanistic account provides with respect to the various norms that govern the construction of explanatory mechanistic models of cognitive capacities. Finally, I highlight the issues which, having been partially analysed in the context of the mechanistic view of cognitive explanation, lead to further important questions that will be discussed in the following chapters of the thesis.

3.2 MECHANISMS AND MECHANISTIC EXPLANATIONS

The core idea of the mechanistic conception of explanation is that a significant class of scientific explanations are a special sort of mechanistic description .[1] More precisely, on this view, scientific models and theories developed in certain areas of scientific inquiry have an explanatory function to the extent and only to the extent that they exhibit the causal mechanisms that maintain, produce, or underlie

1 The mechanistic view of explanation has been developed and defended especially in the context of special sciences such as biology or the different sub-branches of cognitive science (e.g., Woodward 2003; Craver 2007b; Craver and Piccinini 2011). In consequence, most mechanists are willing to admit that fundamental sciences, i.e., the different branches of physics, also use other non-mechanistic styles of explanation (e.g., explanations appealing to symmetry and conservation principles in physics). However, the limitations of the mechanistic view of explanation are less clear when it comes to its application to different branches of the special sciences, such as cognitive science.

the phenomena being investigated. The mechanistic view of explanation has been defended by philosophers of science, such as Bechtel and Richardson (1993/2010), Machamer, Darden, and Craver (2000), Glennan (2002, 2005), Bechtel and Wright (2007, 2009), Bechtel (2008, 2011), and Craver (2007b). Since they consider explicitly the application of the mechanistic framework in the context of cognitive and systems neuroscience, for the purposes of this chapter I will focus on the specific versions of mechanism defended by Craver (ibid.), Kaplan and Craver (2011), and Craver and Piccinini (2011). However, their notion of mechanistic explanation shares many features with other mechanistic accounts proposed in the literature. In what follows, I emphasise two general aspects of the mechanistic conception of cognitive explanation.

Firstly, mechanistic explanation is a form of constitutive explanation that proceeds by decomposition. That is, the behaviour or function performed by a given complex system is analysed and explained in terms of the behaviour or functions of its component parts, their properties, relations, and modes of organisation. There are two important corollaries of this idea. On the one hand, the decompositional strategy implies that the component parts of a mechanism (entities, activities, etc.) are explanatorily prior or more fundamental than the complex mechanism as a whole.[2] This should not be read as claiming that the component entities and activities of a complex system are simple *tout court* or that the methods required to track and model their functions/behaviours are less complex than the ones used to study the target system as a whole. Rather, the explanatory priority of the parts over the whole is meant to reflect the fact that, on the mechanistic view, the component parts and their activities determine the behaviour of the whole system. On the other hand, it is claimed that the mechanistic decomposition of a complex system can be carried out at different levels of organisation or resolution, yielding a hierarchy of (potentially explanatory) mechanistic descriptions.

Secondly, mechanistic explanation is meant to be a form of causal explanation. On the mechanistic view, explanations exhibit the mechanisms that produce the observed behaviour or function of a target system, thereby revealing the relevant causal structure in which the system in question is embedded (cf. Salmon 1989; Craver 2007b). By revealing how the component parts of a mechanism, their activities, interactions, and orchestrated organisation are responsible for a particular observable phenomenon, mechanistic explanations are taken to track genuine causal relations which hold between the component parts of a complex system. That is, according to most versions of the mechanistic view, a causal relation is revealed when a *bona fide* mecha-

2 Psillos (2011, p. 772) makes a similar point in writing that: 'The priority of the parts over the whole - and, in particular, the view that the behaviour of the whole is determined by the behaviour of the parts is the distinctive feature of the broad account of mechanism.'

nism is discovered: 'mechanisms are taken to be the bearers of causal connections. It is in virtue of them that the causes are supposed to produce the effects' (Psillos 2011, p. 773).

3.2.1 *Advantages of the mechanistic conception of explanation*

Some defenders of the mechanistic conception of explanation (e.g., Craver 2007b; Bechtel 2008; Kaplan and Craver 2011; Craver and Piccinini 2011) have argued that the account has two distinctive virtues that recommend it as a general approach to cognitive explanation, to the detriment of alternatives. I briefly characterise these purported advantages before analysing in more detail one specific formulation of the mechanistic criterion for distinguishing explanatory from non-explanatory models/theories of cognitive phenomena.

First, it has been pointed out that mechanistic decompositions may reveal variables (i.e., entities or activities) which can be manipulated and controlled for various experimental and, in some cases, clinical purposes.[3] In fact, the notion of control plays a dual role in the mechanistic conception of explanation. On the one hand, the discovery of mechanisms is supposed to be a sufficient condition for the identification of the features or variables of a given system which can be controlled and intervened upon for different pragmatic purposes (e.g., testing, fixing, repairing, etc.). Thus, mechanistic explanations are said to be desirable because they facilitate the realisation of these sorts of scientific purposes. On the other hand, though, the possibility of controlling certain variables or features of a system is meant to guarantee that the mechanisms that are postulated in the context of specific modelling activities are real and not some artefacts of the experimental and measuring procedures. Thus, the connection between the discovery of underlying mechanisms and control affords a compelling characterisation of the outcomes of certain explanatory activities in the cognitive domain. That is, the mechanistic view shows that explanations sometimes generate new ways of thinking about and intervening upon particular systems for the purpose of bringing about a wide range of outcomes and effects.

Second, the mechanistic view implies that there is an important distinction between merely descriptive phenomenological models and mechanistic models that are viable candidates for explanation. According to mechanists, a phenomenological model is a model that characterises faithfully its *explanandum* and thus can be said to sat-

3 Along these lines, Salmon (1989, p. 812) writes that: '[e]xplanatory knowledge opens up the black boxes of nature to reveal their inner workings. It exhibits the ways in which the things we want to explain come about.' Kaplan and Craver (2011, p. 611) endorse a very similar idea when writing that: 'advances in mechanistic explanation have revealed new knobs and levers in the brain that can be used for the purposes of manipulating how it and its parts behave. That is just what mechanistic explanations do.'

isfy the epistemological requirement of 'saving the phenomena'. It is also acknowledged that some phenomenological models may even license useful predictions about the systems being modelled (the standard example of a phenomenological model with predictive power invoked in the mechanistic literature is the Ptolemaic model of the solar system).[4] However, it is argued that this is not enough to grant explanatory power to phenomenological models, for two reasons (cf. Kaplan and Craver 2011).

The first is that prediction is not a sufficient condition for explanation. For instance, one can predict a storm from the falling mercury in a barometer, but the latter does not in turn explain the storm. The second reason is that phenomenological models are *prima facie* indiscriminate with respect to the kind of details that are deemed explanatorily relevant and those that are not. Inclusion of irrelevant or random detail in the description of a model may impede explanation and mislead future research. Hence, mechanists conclude that phenomenological models, although useful heuristic tools in certain contexts, cannot be *genuinely* explanatory.

However, calling phenomenological models mere heuristic tools neither completely elucidates the difference between mechanistic and phenomenological models nor clarifies what makes the former, but not the latter, *genuinely* explanatory. A more comprehensive account would need to comprise both: (i) a survey of the various epistemic roles played by phenomenological models in scientific research, and (ii) a clarification of the *mechanistic criterion for explanatory relevance*. Whereas a systematic pursuit of the first task has been largely ignored in the mechanistic literature, proponents of the mechanistic account usually acknowledge the importance of articulating a precise criterion for the explanatory relevance of the factors cited by particular mechanistic models.

3.2.2 *Mechanistic explanatory relevance*

The problem of explanatory relevance confronting the mechanistic view consists in finding a principled way to determine which of the features of a mechanistic model are explanatorily relevant and which are not. In a recent formulation of the mechanistic criterion for explanatory relevance, Kaplan and Craver (ibid., p. 611) hold that an explanatory model has to satisfy a model-to-mechanism-mapping (3M) constraint:[5]

4 More relevantly, most mathematical models developed in cognitive and systems neuroscience are also taken to qualify as mere phenomenological models which 'summarise large amounts of experimental data compactly yet accurately, thereby characterising what neurones or neural circuits do' (Dayan and Abbott 2005, p. xiii).

5 Strictly speaking, the 3M constraint requires only that the entities and relations postulated in a scientific model be interpretable in a concrete (physical) vocabulary. In order for the 3M constraint to count as a proper mapping account (i.e., in the

> (**3M**) In successful explanatory models in cognitive and systems neuroscience (a) the variables in the model correspond to components, activities, properties, and organisational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (b) the (perhaps mathematical) dependencies posited among these variables in the model correspond to the (perhaps quantifiable) causal relations among the components of the target mechanism.

The 3M requirement summarises the principal ingredients of a mechanistic view of explanation. First, there is the commitment to the idea that mechanisms are the crucial explanatory tools used in cognitive and systems neuroscience.[6] Second, 3M includes a strong representationalist criterion according to which abstract (mathematical) relations holding between the mechanistic variables of a system must correspond to causal relations among the component parts of the modelled mechanism. In other words, 3M expresses a twofold commitment to (i) the idea that only mechanistic models of cognitive capacities count as 'successful' explanations and (ii) the notion that the systems investigated by cognitive science are, in an important (broadly metaphysical) sense, mechanisms.

However, this picture of mechanistic explanation is not entirely complete without the idea of a hierarchy of mechanistic decompositions mentioned above. With respect to this, Craver (2006b, p. 360), writes that: '[m]odels that describe mechanisms can lie anywhere on a continuum between a *mechanism sketch* and an *ideally complete description of the mechanism*.' In other words, mechanistic decompositions can be offered at different levels of organisation and they span a continuum of decompositional descriptions ranging from mechanistic sketches, mechanism *schemata*, up to ideally complete mechanistic descriptions.[7]

The crucial difference between the members and their position on this scale of mechanistic models consists in the amount of detail that

model-theoretic sense), mechanists should specify the two structures that are to be connected via a particular type of mapping. However, in its current formulation, the 3M account requires only that the weaker interpretability condition be met.

6 Although they provisionally restrict 3M to the field of cognitive and systems neuroscience, Kaplan and Craver (2011, 610, ftn.10) acknowledge that they 'see no good reason to exempt all of cognitive science from the explanatory demands laid out by 3M'; see also Craver and Piccinini (2011). In light of these (implicit) commitments, the argument of the paper will continue to refer to the application of the mechanistic view to cognitive science, broadly conceived (see chapter 2).

7 'A mechanism sketch is an incomplete model of a mechanism. It characterises some parts, activities, and features of the mechanism's organisation, but it has gaps'; while 'ideally complete descriptions of a mechanism [...] include all of the entities, properties, activities, and organisational features that are relevant to every aspect of the phenomenon to be explained.' In between, he says there is 'a continuum of mechanism schemata that abstract away to a greater or lesser extent from the gory details [...] of any particular mechanism' (Craver 2006b, p. 360).

each of them presupposes. For instance, mechanism sketches, which are usually specified only at the functional level, are said to include many filler terms that require further completion and specification (cf. Craver 2007b; Craver and Piccinini 2011). Mechanism *schemata* are more specific and include extra detail about the component parts of a mechanism, but are still incomplete and do not qualify as genuine mechanistic explanations.

A last conceptual distinction that mechanists introduce in order to clarify the difference between explanatory and non-explanatory models is that between *how-possibly* mechanisms and *how-actually* mechanisms (cf. Craver 2006b, 2007b). Whilst there is some overlap between this latter distinction and the categories of mechanism sketches and *schemata*, only *how-actually* mechanistic models count as genuinely explanatory.[8] *How-possibly* mechanisms, alongside mechanism *schemata* and mechanism sketches, may serve as useful tools that aid in the search for explanation, but they do not have an explanatory function because they do not represent the real (actual) mechanisms underlying the target phenomena.[9]

It is reasonable therefore to conclude that, according to this formulation of the mechanistic view, scientific models of cognitive capacities are explanatory to the extent and only to the extent that they exhibit *how-actually* mechanisms. Moreover, given the explanatory priority of the component parts of complex systems and the notion of hierarchical mechanisms, there is a *prima facie* case for believing that mechanistic explanation works essentially in a bottom-up fashion, by specifying the lower-level mechanisms supporting the target higher-level cognitive phenomena. The latter claim concerns the specific type of *explanatory structure* associated with mechanism, but might not necessarily be appropriate for characterising other features of the mechanistic theorising and experimental practices.

Nevertheless, as will become more obvious in the following sections, this entailment of the mechanistic view seems to clash with the idea that the mechanistic framework actually facilitates a series of inter- and intra-level interfield integrations in the area of cognitive science (Craver 2007). I aim to clarify the sources of this potential tension which lies at the heart of the mechanistic view, while defending its valuable insights concerning the integration of particular hypotheses developed at different levels of analysis of cognitive phenomena.

8 'How-actually models describe real components, activities, and organisational features of the mechanism that in fact produces the phenomenon. They show how a mechanism works, not merely how it might work.' (Craver 2006b, p. 361)

9 'How-possibly models (unlike merely phenomenal models) are purported to explain, but they are only loosely constrained conjectures about the mechanism that produces the explanandum phenomenon. They describe how a set of parts and activities might be organised such that they exhibit the explanandum phenomenon.[...] How-possibly models are often heuristically useful in constructing a space of possible mechanisms, but they are not adequate explanations.' (ibid.)

In the next section, I turn to analysing some of the most prominent challenges that face the mechanistic model of cognitive explanation.

3.3 THE LIMITS OF MECHANISM

This section challenges what I take to be an ambitious presupposition underlying the applicability of the mechanistic conception of explanation to cognitive science. More precisely, I argue that there is a strong tendency in the mechanistic view of cognitive explanation to claim that mechanism is the only *genuinely* explanatory framework appropriate for the study of cognitive phenomena. This tendency is transparent in two particular argument strategies utilised by mechanists. First, there is the 'what-else' argument, according to which no other framework except mechanism is able to provide a descriptively adequate and normative picture of cognitive explanation (Kaplan and Craver 2011; Craver 2007b). Second, mechanists also rely on an inductive strategy by arguing that since mechanism has proven to be a 'successful' explanatory strategy in certain branches of cognitive science, then it must be possible to extend it to other domains as well (Craver and Piccinini 2011). Together, the two types of arguments appear to license, from a 'local' point of view, a general thesis about the wide scope of mechanistic explanation within cognitive science.

3.3.1 *Ontic mechanistic explanations*

For the purpose of the following argument, I propose to distinguish between: (i) a strong realist strategy, (ii) a moderate realist strategy, and (iii) an epistemic strategy of defending the mechanistic view of cognitive explanation. Although there is some variation in the ways in which different mechanists present their position, I claim that most of their arguments can be grouped under one of these three strategies. My contention will be that none can support the bold presupposition of the mechanistic account, viz. that mechanism provides the *only* genuine explanatory framework appropriate for cognitive science. However, I also show how one of these strategies can be interpreted so that it reflects the important contributions of mechanism to the study of certain aspects of cognitive phenomena.

The strong realist strategy is implicitly at work in some of the arguments put forward by certain mechanist philosophers (e.g., Craver 2007b, 2012; Strevens 2008) and presupposes a broad metaphysical picture of *real* mechanisms. On this account, any complex entity or structure whose function is determined by the organised functioning of its component parts and activities counts as a mechanism.[10]

10 The standard characterisation of mechanisms is due to Machamer et al. 2000, who write: 'Mechanisms are entities and activities organised such that they are productive of regular changes from start-up to finish or termination conditions' (3). For al-

Thus, the metaphysical picture implicit in the mechanistic conception is seemingly very plausible: it requires only a realist commitment to entities and their activities that are organised together in order to yield complex mechanisms whose overt behaviour they explain.

However, there seem to be two straightforward problems with this 'thin' mechanistic metaphysics. The first arises from an ambiguity concerning the metaphysical status of mechanisms themselves. That is, it is not entirely clear whether by 'real' mechanisms mechanists mean *particular* or *universal* mechanisms. If the former, then it follows that mechanistic explanation is always explanation of a particular phenomenon/event.[11] If the latter, then the strong realist commitment towards universal mechanisms would be a form of realism about abstracta which would not fit well with the mechanistic insistence that mechanisms are things that can be manipulated and intervened upon. This in turn further reinforces the idea that mechanistic decomposition is primarily a tool for explaining particular occurrences of physical phenomena/events. Either way, it seems that mechanists need to be more specific about which conception they take to be correct before claiming that *real* mechanisms are the ones doing the explanatory work in a scientific model or theory.

The second problem concerns the metaphysical consequences of taking mechanisms to be the building blocks of nature (and the fundamental constituents of causation). On the assumption that mechanisms are the building blocks of nature, the mechanist seems to be forced into one of the following two options. Either mechanisms go all the way down and there is no fundamental level of mechanisms that constitute the bare bones of natural phenomena, or there is a fundamental mechanistic level at which explanation also stops and is deemed to be complete. The first horn of the *fundamental level* dilemma is arguably problematic because it presupposes that there is no metaphysical basis of the hierarchy of mechanistic models. This in turn seems to imply that there is no determinate way to establish what mechanisms are real and explanatory with respect to the target physical phenomena. The second horn of the dilemma assumes that there is an ultimate mechanistic level to be discovered, presumably by a future fundamental physics. Beside its being questionable whether fundamental physics is in the business of discovering the mechanistic structure of the world (cf. Schaffer 2003; Ladyman and Brown 2009; McKenzie 2011), adopting this horn of the dilemma would still undermine the mechanistic explanatory relevance criterion simply because

ternative accounts which allow feedback loops and other forms of self-organisation of the component parts and activities of a mechanism, see Abrahamsen and Bechtel 2006; Bechtel and Abrahamsen 2010, 2013.

11 Salmon (1989, p. 184), for instance explicitly says that mechanistic explanations are local, 'in the sense that they show us how particular occurrences come about; they explain *particular* phenomena in terms of collections of *particular* causal processes and interactions.' (m.i)

we do not yet have the mechanistic criteria provided by this future fundamental science.[12]

In consequence, on both horns of the *fundamental level* dilemma, one is left with an indeterminacy issue about *how* to determine the *real* mechanisms that explain the observable cognitive phenomena. Otherwise put, since neither solution to the fundamental level dilemma is metaphysically unproblematic, mechanists have very good reasons to avoid the strong realist strategy in promoting a mechanistic view of cognitive explanation. I now turn to examining whether their moderate realist strategy fares any better.

The moderate realist strategy is perhaps the most attractive mechanistic position. It starts from the description of the explanatory practices of an important research community within cognitive science (e.g., molecular neurobiologists) and it extracts from it a general picture of the scientific explanation of cognitive phenomena. This picture exploits two widely held assumptions about the structure of explanation that depend crucially on the fact that most explanations target complex systems. The first one is that explanations of complex systems should be constitutive, i.e., they should exhibit the parts that compose the complex system. The second assumption is that the behaviour of a complex system will be intelligible in terms of the behaviour of its component parts. Both assumptions express the intuitive epistemological requirement that explanations make things perspicuous by breaking them into smaller, more tractable pieces.

Unlike the previous strategy, the moderate realist strategy points out that practising cognitive scientists and neuroscientists are in the business of discovering the real mechanisms that underlie the observable phenomena they want to explain. Although talk of 'discovering real mechanism' may still be taken to have some metaphysical undertones, mechanists often insist that the term 'real mechanism' merely stands for the kinds of things discovered, manipulated, and controlled by practicing scientists. Moreover, they claim that taking these experimental and theoretical practices at face value blocks any serious skeptical arguments concerning the reality of mechanisms posited in the course of scientific inquiry.

Despite the appeal of this move, there still seems to be a *prima facie* tension between the moderate realist's implicit unconditional reliance on current scientific practice and the stricter criterion for mechanistic

12 It should be noted that the argument proposed against the mechanist's strong realist strategy does not depend on any particular stance that one might take with respect to fundamentality questions in the physical sciences. The point being made is only that under a strong realist reading of the mechanistic thesis one has a problem determining which mechanisms are genuinely explanatory in the first place. However, more generally, I take it that fundamentality questions (like any other ontological questions) ought not to be settled *a priori* but should be a matter, where possible, of empirical enquiry (cf. McKenzie 2011). This is in line with the underlying assumption of the thesis that ontological questions/principles become operative when taken in the context of particular (scientific) practices.

explanatory relevance captured by the 3M constraint. The fact that the mechanistic models developed in scientific practice are empirical hypotheses means that they are falsifiable, which in turn implies that a mechanistic model may be deemed to be explanatory at a certain time but not later. Otherwise put, the exclusive reliance on scientific practice does not guarantee that there is a decisive way of determining *once and for all* the explanatory power of a particular mechanistic model.

In addition, there is another issue which indicates that adopting a practice-based perspective might not suffice to solve the general problem of the explanatory relevance of the mechanisms discovered in the course of scientific inquiry. Molecular biologists might discover the *real* mechanisms underpinning the formation and propagation of action potentials in nerve cells (cf. the Hodgkin-Huxley (1952) model discussed in Craver 2006b, 2007b), but these in turn might be decomposed, and thereby explained, in terms of simpler biochemical mechanisms. The apparent problem follows from the two requirements implicit in the mechanistic model of explanation, viz. that explanation proceed in a bottom-up fashion and that only the more detailed mechanistic descriptions be deemed genuinely explanatory. For in light of these requirements, only the lower-level mechanistic model (i.e., the biochemical one) seems to count as genuinely explanatory. However, this conflicts with the mechanistic analyses of molecular approaches in cognitive science (Craver 2007b; Bechtel 2008; Bechtel and Richardson 1993/2010), according to which molecular mechanisms can also be taken to explain a range of appropriately circumscribed cognitive phenomena.

As will be shown below, there is a way of avoiding both of these issues by appealing to the idea that mechanistic explanations are in fact multilevel integrated accounts of the phenomena they are supposed to model. Before considering how the notion of piecemeal intra- and inter-level integration is supposed to work (in principle) in a domain such as cognitive science, I will consider another strategy for promoting the mechanistic conception of cognitive explanation which offers further clues about how to interpret, in the most profitable way possible, the mechanistic view of explanation.

3.3.2 *Epistemic mechanistic explanation*

Whereas proponents of the realist strategies hold that mechanistic explanation essentially involves fitting a phenomenon into the causal structure of the world (e.g., Salmon 1984a; Craver 2007b, 2012), defenders of the epistemic strategy (e.g., Bechtel 2008; Wright 2012) emphasise the fact that explanation is a human activity which is primarily concerned with understanding the phenomena being investigated.

For instance, Bechtel (2008, p. 18) explicitly states that: '[e]xplanation is fundamentally an epistemic activity performed by scientists.'[13]

There are two important corollaries of the epistemic strategy of defending mechanism. Firstly, the epistemic view entails that the construction and evaluation of good mechanistic explanations of specific cognitive phenomena depends to a significant extent on the aims and purposes of particular research communities (e.g., molecular neurobiologists, computational neurobiologists and cognitive scientists, etc.). This seems to provide one potential solution to the puzzle raised above concerning the explanatory power of alternative mechanistic models. For the evaluation of the explanatory value of specific mechanistic models would have to take into account the distinct aims pursued by different research communities which develop mechanistic models/descriptions at different levels of abstraction. Secondly, under the view that mechanistic explanation is a complex epistemic activity, pursued by many scientists, it has been argued that mathematical objects should be treated as part of mechanistic explanations proper (Abrahamsen and Bechtel 2006; Bechtel and Wright 2007). Otherwise put, since these abstract tools are required in order to make tractable complex cognitive phenomena, defenders of the epistemic strategy have insisted that they should be viewed as an integrative part of *dynamic mechanistic* explanations (Bechtel and Abrahamsen 2010, 2013; Bechtel 2011).

In response, proponents of the realist or ontic view of mechanistic explanation have argued that the epistemic strategy makes mechanistic explanations depend too much on the interests and goals of particular research communities, which in turn undermines their objectivity (e.g., Craver 2012). Realists claim that the epistemic strategy transforms mechanism into a relativist position that assigns an explanatory role to almost any mechanistic model that satisfies some set of suitable epistemic constraints. In addition, realists insist that mathematical tools play only a heuristic role in circumscribing and/or describing the cognitive patterns to be explained by the causal mechanisms exhibited by *how-actually* mechanistic models. The implicit assumption underlying this argument seems to be that the main strength of explanatory mechanism within the cognitive domain consists in

13 The contrast between the two conceptions can be made more acute by pointing out that, on the ontic conception, mechanistic explanations are 'objective portion[s] of the causal structure of the world, [...] the set of factors that bring about or sustain a phenomenon.' Or, 'objective explanations are not texts; they are full-bodied things. They are facts, not representations' (cf. Craver 2007b, p. 27). On the other side, for proponents of the epistemic conception, mechanistic explanations are texts, or descriptions that aim to increase our knowledge of observable phenomena. They write that: '[o]bviously, knowledge of how things work is an epistemic matter if anything is, which is just to say that analysis of mechanistic explanatory texts properly requires a broadly epistemic conception of mechanistic explanation' (cf. Wright 2012, p. 382).

its commitment to the *existence* of certain (types of) neurobiological mechanisms.

My contention is that the moderate realist strategy need not, and indeed should not, confine the explanatory value of mechanistic models exclusively to their biological realist commitments. The main insight of the moderate realist is that ontic principles play a crucial role in the development of (good) mechanistic explanations. Still, this need not exclude the possibility that other epistemic constraints or principles might play an equally important role in constructing and evaluating mechanistic models of particular cognitive phenomena. In fact, granting that the construction and evaluation of particular explanations of cognitive phenomena is a complex epistemic activity, ontological principles can be viewed as things that must be assumed in order for the explanatory practices to get off the ground. As such, ontological principles reflect the way in which explanation is connected to other scientific activities such as measurement, control, and confirmation, that complement the scientific investigation of cognitive phenomena.

In a nutshell, the proposal is to conceive the moderate realist strategy as a way of emphasising the norms and principles that guide the practice of building and evaluating mechanistic models of specific cognitive capacities. These include both ontic commitments to the parts and activities of complex mechanisms detected with the help of existing experimental tools and techniques, as well as epistemic principles reflecting the knowledge, interests, and aims of different groups of researchers. As Phyllis Illari (2013) has pointed out, this reinterpretation of mechanism helps to resolve the apparent tension between realist/ontic and epistemic conceptions of mechanism, while preserving the core structure of mechanistic explanation.

Another important consequence of the proposed reconstruction of the moderate realist view of mechanistic explanation is that it shows that the normative constraints identified by mechanist philosophers are essentially tied to specific systems of scientific practice (i.e., theories, research programmes). For this reason, they cannot be taken to determine the necessary and sufficient conditions for something to count as a good explanation of a particular cognitive phenomenon *tout court*. This casts some doubt on the claim that mechanism can offer a general unifying framework for the different styles of explanation deployed in cognitive science. Therefore, contra the strong mechanistic contention, I claim that the explanatory value of a particular scientific theory/model ought not to be equated to a set of existential claims about the mechanisms (their component parts, properties, and activities) underlying a particular cognitive capacity. Instead, the reinterpreted moderate realist strategy which stresses the normative role played by specific ontological principles provides a better, non-polarised philosophical perspective on the practice of constructing

and evaluating a particular style of explanatory models of cognitive capacities.

The above arguments suggest a shift of emphasis away from the idea that mechanism can offer a general account of cognitive explanation and towards the analysis of the different scientific activities that yield potentially explanatory models of specific cognitive phenomena. This move implies that mechanistic models offer only partial accounts of the cognitive phenomena they are intended to explain. However, this partiality does not derive from the notion that a complete explanatory theory of cognition is only an ideal that guides scientific inquiry (cf. Railton 1989; Craver 2007b). Rather, I take the partiality of mechanistic explanation to follow from recognising that there are other proper explanatory strategies/schemas that can be used in the investigation of various aspects of cognition.

I turn next to a paradigmatic example of scientific modelling used in neurocognitive research (the Difference-of-Gaussians (DOG) Model of visual spatial receptive field organisation) in order to show how the explanatory value of abstract models can also be justified independently of mechanism in light of current neuroscientific practice.

3.4 NON-MECHANISTIC MODELS REVISITED

It has been argued above that the explanatory power of mechanistic models/theories consists in the specification of the organised interaction of the component parts and activities of the mechanisms that underlie the cognitive phenomena being modelled. As such, mechanism has been taken to challenge the claim that dynamical system theory, with its focus on mathematical models of cognitive phenomena, can provide an adequate account of cognitive explanation (Chemero and Silberstein 2008; Stepp, Chemero, and Turvey 2011). In what follows, I propose to draw on the analysis of the DOG model to articulate a series of important lessons about the roles of mathematisation in the study of cognitive phenomena which will also shed light on the explanatory claims of dynamical systems theorists. The model analysed below is part of an important area of research in vision studies that focuses on building robust models of receptive field structures that can explain neuronal responses to different classes of visual stimuli.

3.4.1 *The Difference-of-Gaussians Model of Visual Spatial Receptive Field Organization*

One general aim of vision studies is to offer an account of how the visual system extracts structure out of the information received from the external environment. In particular, it aims to provide an account of how the visual system is capable of performing tasks such as edge detection, shape, or colour perception. In order to say how the visual

system performs these complex tasks, researchers study more basic properties of components of the visual system, like the selectivity of each early visual neurone to movement, line orientation, and other features of the visual stimuli. The selectivity of each of these neural cells to such parameters is in turn determined to a great extent by the structure of the neurone's receptive field. Roughly put, a cell's receptive field is a restricted region of visual space where a luminous stimulus could change the cell's level of activation.

The DOG model was proposed by Rodieck (1965) to characterise the structure of the receptive fields of a class of neurones which play an important role in early vision, namely ganglion cells. These cells gather signals from multiple photoreceptors and transform them into trains of action potentials which in turn propagate to various areas in the visual cortex. A distinctive feature of ganglion cells is that they have small receptive fields with a simple organisation, which resembles two concentric circles, usually known as the centre-surround organisation. This organisation is primarily responsible for the selectivity of ganglion cells to certain features of the visual stimuli. There are two types of ganglion cells: ON-centre retinal ganglion cells respond to light spots surrounded by dark backgrounds, whereas OFF-centre retinal ganglion cells respond to dark spots surrounded by light backgrounds. As a first step, Kuffler (1953) modelled the receptive fields of ganglion cells as two distinct, concentric, and mutually antagonistic regions. Then, Rodieck (1965) characterised their structure as the arithmetic difference between two Gaussian functions, as follows:

$$F(x,y) = \frac{A_1}{2\pi\sigma_1^2}\, exp(-\frac{x^2+y^2}{2\sigma_1^2}) - \frac{A_2}{2\pi\sigma_2^2}\, exp(-\frac{x^2+y^2}{2\sigma_2^2}),$$

The two terms of this function represent the ON-centre component of the response (the relative change in firing rate after stimulation, also known as the sensitivity distribution) and the opposite-signed OFF-centre surrounding component. The coefficients A_1 and A_2 mark the peak local sensitivity of the centre and of the surround, respectively, while σ_1 and σ_2 correspond to the width parameters of the two Gaussian envelopes of the centre and surround.

A series of features of the DOG model are relevant for assessing its scope and explanatory function. Firstly, Rodieck (ibid.) had not constrained his mathematical description so that the variables in the DOG model may be mapped to identifiable components, organisational features, and operations of the synaptic mechanisms producing the spatial organisation of the cells' receptive fields. Partly because of this, it was possible to show that the model is compatible with a series of (how-possibly) underlying mechanisms (e.g., Cohen and Sterling 1991; Einevoll and Heggelund 2000). Also, due to the same sort of mechanistic indeterminacy, the model cannot be taken to provide a complete explanatory account of *why* ganglion cells respond in the way that they do to retinal inputs. That is, the mathematical model

does not provide any insight or information about how synaptic or molecular mechanisms contribute to the ganglion cells' selectivity to specific features of the visual stimuli.

However, on the positive side, the model exhibits the relevant parameters on which the spatial organisation of the retinal ganglion cells' receptive fields depends, viz. the shape of the sensitivity distributions for the two main regions of the cell. In doing so, it rules out a variety of other parameters which would appear in a complete description of the specific response of ganglion cells to retinal inputs, viz. their shape, size, and the number of corresponding photoreceptors. The model also characterises the specific dependence relation between the response profiles of the two components of the spatial receptive field, viz. the centre and the surrounding. This dependence relation is the dominant feature of the receptive field structure in virtue of which one can then account for the observed selectivity of retinal ganglion cells to particular features of the visual stimuli.

Moreover, the DOG model has been tested on a variety of different stimuli, including circles, moving bars, and sinusoidal gratings (Einevoll and Plesser 2005) and it has been extended to characterise the spatial field organisation of certain visual cells in dLGN (Dawis and Tranchina 1984). As such, the model can be taken to identify a very general feature of certain classes of neural cells in the retina and the thalamus in virtue of which these cells are responding selectively to particular features of the visual stimuli.

Despite its capacity to characterise accurately the target pattern in early visual processing - viz., the specific sensitivity distributions of retinal ganglion cells - and to license new predictions about the profile responses of other types of neural cells, mechanists claim that the DOG model does not play any explanatory role in vision studies (cf. Kaplan and Craver 2011). *Contra* dynamicists, they claim that the features emphasised above (empirical adequacy, unificatory, and predictive power) do not suffice to make a cognitive model genuinely explanatory.[14] In addition, they claim that the model should satisfy the standards expressed in the model-to-mechanism-mapping (3M) constraint. In order to strengthen their point, Kaplan and Craver (2011) claim that the mathematical DOG model can be turned into a *genuinely explanatory mechanistic model* by adding information about the anatomical connections between retinal ganglion cells and the photoreceptors whose signals they are sensitive to. They state that: '[t]he explanatory step will clearly involve understanding neuronal morphology and the synaptic connections in the retina and perhaps

14 For instance, Stepp, Chemero, and Turvey (2011, p. 435) characterise the explanatory function of dynamic models precisely along these lines when they write that: 'dynamical systems models provide law-like explanations, support counterfactuals, and allow predictions that can be used to guide future experimental research; the best dynamical models can be used to unify disparate phenomena, capturing them under a single explanatory scheme.'

the developmental processes by which such functional relations are constructed, elaborated, and maintained' (Kaplan and Craver 2011, p. 621).

3.4.2 *Mathematical models and explanatory structure*

I agree with Kaplan and Craver (ibid.) that current formulations of the dynamicist account of cognitive explanation run the risk of being conflated with predictivism, i.e., the view that the explanatory value of a model/theory derives from its descriptive and predictive power. Nevertheless, I believe that there is a way to vindicate the dynamicist intuition according to which certain abstract models explain particular aspects of their target cognitive phenomena without invoking the causal mechanisms that produce or support them. In a nutshell, the proposal is that abstract (including mathematical) models are explanatory if they reveal certain features of the target cognitive patterns/phenomena which count as being *more fundamental* than the ones identified in the description of the explanandum. In addition, one should be able to specify a number of steps through which the explanandum is linked to the more fundamental structure specified by the explanans. Although a more precise characterisation of the notion of fundamentality implicit in this conception of explanation will have to be postponed for the last chapter of the thesis, a preliminary definition should suffice to clarify the present point.

On the view I put forward, the structure postulated by the explanans of a particular explanation is more fundamental than the explanandum if it can be used to support a set of relevant counterfactual generalisations pertaining to the phenomenon being investigated, and generate further structures/concepts which yield insight and understanding of that phenomenon. Whilst what is fundamental is standardly defined as that which is ontologically dependent on nothing else, on the proposed view something counts as being more fundamental only against the background of a particular system of knowledge (embodied in a scientific theory and/or experimental practice). This conception of fundamentality is intended to reflect the idea that ontological dependence relations/principles become operative only when considered in the context of specific scientific practices. Thus, in different investigative contexts, different ontological principles might turn out to be relevant for determining what counts as the more fundamental structure of the system/phenomenon being investigated. Moreover, this conception implies that there is no way of determining either *a priori* or in an absolute sense what counts as the more fundamental, and thereby the explanatory structure that would help elucidate a given target phenomenon.

Now, if the concepts/structures used to characterise, in a particular stepwise fashion, certain fundamental features of a target cogni-

tive phenomenon have a distinctive mathematical character, then that cognitive explanation is adequately characterised as a mathematical explanation. The mathematical explanation would thus show a way to connect the description of the cognitive phenomenon provided by the explanandum with the fundamental feature identified and characterised by the mathematical model. In other words, an explanatory mathematical model is one that shows that certain features of a cognitive system (specified by the explanandum) hold in virtue of other more fundamental (mathematical) features that can be determined independently from the specification of any particular causal laws or mechanisms.

This latter characterisation of the source of the explanatory power of cognitive mathematical models also allows us to appreciate the sense in which mathematical (e.g., dynamic systems) models are non-causal forms of explanation. These models are non-causal because they do not work by providing information about a given cognitive phenomenon's causal history, or about the causal mechanisms through which the modelled phenomenon is embedded in a larger network of causal relations. As pointed out previously, the original DOG model (Rodieck 1965) had not been constrained so as to reflect the causal mechanisms underlying the selective response of ganglion cells to multiple photoreceptors. This and other mathematical models are potentially explanatory in virtue of the fact that they show that the cognitive pattern to be explained holds in virtue of a feature which does not correspond to any specific causal property or process but rather is more fundamental than the features exhibited by the underlying causal mechanisms. The advantage of this characterisation of the potential explanatory value of mathematical models of cognitive phenomena is that it is compatible with the features of dynamical systems models discussed in the literature, namely the capacity to support counterfactual generalisations, to guide experimental research, and to unify apparently disparate phenomena (cf.Stepp, Chemero, and Turvey 2011).

However, the proposed formulation constitutes an extension of existing accounts insofar as it does not equate the explanatory value of mathematical models either with their predictive or their unificatory power. Rather, it highlights the fact that cognitive explanations that make an essential appeal to mathematical concepts are one possible way of structuring our knowledge of cognitive phenomena which affords a better understanding of some of their features (just like mechanistic decompositions are yet another way of achieving a similar goal, as will be argued below). In light of these considerations, I claim that mathematics does sometimes play a genuine explanatory role in cognitive modelling.

Nevertheless, claiming that certain uses of mathematics in cognitive modelling have an explanatory value does not necessarily rule

out there being other aspects of mathematisation that are consistent with the mechanistic model of explanation. For instance, Kaplan and Craver (2011, p. 605) point out that 'mechanisms are often described using equations that represent how the values of the component variables change with one another.' They take such mathematical descriptions to be useful tools for 'characterising complex interactions among components in even moderately complicated mechanisms.' Bechtel (2011) also recognises that mathematical tools are in some cases indispensable for making tractable certain complex cognitive phenomena, and that, for this reason they should be treated as proper parts of dynamic mechanistic explanations. Note that although Bechtel seems inclined to admit the explanatory contributions of mathematical concepts to the understanding of empirical (cognitive) phenomena, he does not provide a separate analysis of how these mathematical structures get to play these explanatory roles in the first place. Moreover, his insistence that mathematical structures should be 'integrated' in mechanistic explanations of cognitive phenomena brings him closer to Kaplan and Craver's (2011) contention that mathematics plays a merely heuristic role in the development of genuinely explanatory mechanistic models.

Since an important target of this chapter has been to identify and criticise the ambitious assumption that mechanism provides the single most appropriate framework for explaining cognition, I now consider briefly the main motivation that seems to underpin the mechanistic resistance to acknowledging the explanatory role of certain abstract (e.g., dynamical) models of cognition. In short, it seems that the standard realist construal of mechanistic explanation (Craver 2007b, 2012) introduces an artificially strong distinction among the notions and tools used in day-to-day scientific explanative activities. More specifically, it generates a strong dichotomy between abstract mathematical structures used as tools in scientific modelling and other theoretical posits such as complex entities, activities, and forms of organisation of purported mechanisms. Tacitly endorsing this metaphysical polarisation, mechanists deny the potential explanatory value of mathematical models, insisting that it would require a problematic ontological commitment towards the abstract structures posited by particular mathematical models. According to the realist, the explanatory value of mathematical scientific models would be secured only via a Platonist commitment to the abstract (mathematical) entities posited by the models in question.

It should be clear from the arguments provided so far that an advocate of the explanatory value of mathematical models of cognition need not be a realist about mathematical or other abstract entities. If the explanatory function of a mathematical model consists in showing how a given cognitive pattern arises in virtue of a (more) fundamental non-causal feature of the system under investigation, then

one should not be forced to pursue the justification of the explanatory value of that particular model beyond this step. This is essentially because, on the proposed conception, the existence or non-existence of the mathematical objects used to capture a fundamental feature of the (cognitive) system being investigated does not add anything to the explanatory structure of the mathematical model.

This line of reasoning reflects an important feature of the account of cognitive explanation that emerges from the present investigation, namely that there is no sharp categorial division between the different tools and concepts used by scientists in developing good explanatory models of concrete (physical) phenomena. If explanation essentially involves exhibiting, unfolding, and connecting salient features/patterns in the functioning of complex physical systems, then it should not be contentious that these regularities might be captured (sometimes exclusively) at specific levels of abstraction (e.g., mathematical). For instance, in the case of the DOG model, the *explanans* consists in a specific, highly general pattern that emerges in the organisation and functioning of retinal ganglion cells. The postulated pattern explains the specific selectivity of this class of neuronal cells to certain features of the visual stimuli.

Assuming that the proposal sketched in this section is on the right track, I take it to indicate a potentially fruitful strategy for mitigating the claim that mechanism constitutes the *single* appropriate explanatory framework for cognitive science. More precisely, this strategy is based on two ideas which have played an essential role in the arguments developed so far. First, I have insisted that the explanatory value of a model depends on whether it succeeds in creating a link between the characterisation of a particular cognitive phenomenon (the explanandum) and the characterisation of a more fundamental aspect of that target phenomenon (the explanans). Thus, on the proposed account, cognitive explanations create a specific type of connection between different ways of conceiving certain aspects/features of a given cognitive phenomenon. From this 'epistemicised' view of cognitive explanation, it follows that existential or realist commitments ought not to be equated with the explanatory value of particular models/theories of cognition.

Second, conceiving of explanation as an epistemic activity rather than as a metaphysical link connecting different types of entities or structures in the world allows us to appreciate that the practice of constructing (good) scientific explanations of cognitive phenomena involves a network of scientific activities, which can vary with the type of explanatory structure proposed by a particular scientific model/theory. Thus, in the case of mechanistic models of cognitive capacities, the relevant epistemic activities which contribute to the construction of good mechanistic explanations include various manipulation, control, and confirmation techniques which increase the

evidential support for the entities and activities postulated by specific mechanistic models. However, potential mathematical explanations of cognitive patterns/phenomena need not be connected in the same way or with the same type of epistemic activities involved in validating mechanistic explanations of cognitive capacities. Nor is the relation between mathematical and mechanistic explanations of cognitive capacities always one of competition. As will be argued below, the proposed picture of the roles of mathematisation in cognitive research, albeit sketchy, helps to solve some of the inconsistencies which threaten the mechanistic picture of cognitive explanation.

3.5 MECHANISMS AND MORE

As stated in the introduction, the mechanistic view of explanation has a series of appealing features which seem to recommend it as a general philosophical picture of the explanatory strategies used across the various sub-branches of cognitive science (cf., Craver 2007b; Craver and Piccinini 2011). In this final section, I revisit these advantages in light of the critical analysis developed above and respond to some additional objections which take into account the normative character and integrative potential of the mechanistic account of cognitive explanation. In replying to these concerns, I aim to provide additional support and motivation for: (i) the argument strategy adopted in this chapter, and (ii) a more cautious endorsement of mechanism as one of the multiple explanatory frameworks used in the study of cognitive phenomena.

3.5.1 *Final objections and replies*

The distinctive feature of mechanistic explanations of cognitive capacities is that they are a type of constitutive de-compositional explanation which accounts for the behaviours of (that is, functions performed by) complex cognitive systems in terms of their parts, their properties, and their organised interactions (e.g., Craver 2007b; Bechtel 2008; Kaplan 2011; Kaplan and Craver 2011; Craver and Piccinini 2011). Section 3 distinguished between three major strategies for defending a mechanistic view of cognitive explanation: (i) a strong realist strategy, (ii) a moderate realist strategy, and (iii) an epistemic strategy. I have shown that the strong realist strategy (e.g., Craver 2006b, 2007b, 2012) confronts a series of difficult problems primarily because it equates the explanatory value of mechanistic models/theories of particular cognitive phenomena to the existential claims about the mechanisms which underlie those cognitive phenomena. Whilst the epistemic strategy (Bechtel and Richardson 1993/2010; Bechtel 2008, 2011) seems to provide a more adequate perspective on the scientific practices of constructing explanatory mechanistic models/theories of

cognitive phenomena, it has nevertheless been criticised for being too permissive with respect to the factors that count as having a genuine explanatory function in such modelling contexts. The analysis of these two strategies led to an interpretation of the moderate realist strategy which represents it as the most compelling way of defending the explanatory contributions of the mechanistic framework to the study of cognitive phenomena.

The proposed interpretation aims to vindicate the intuition that part of the appeal of the mechanistic view of cognitive explanation resides in its biological realist commitments. Most mechanists stress the fact that the explanatory value of mechanistic models of cognitive capacities is strongly related to their biological plausibility. For this reason, they hold that cognitive models/theories which do not make any direct claim concerning the biological plausibility of the theoretical entities and relations that they postulate in accounting for different aspects of cognitive processing are not *genuinely* explanatory. In response, I have argued that there are good reasons to doubt that the biological plausibility (or existence) criterion *per se* will guarantee the explanatory value of particular cognitive theories/models, whether mechanist or not. This is primarily because the proposed models/theories have to confront: (i) the incompleteness of our current biological knowledge, and (ii) the empirical underdetermination of any model/theory of a complex empirical (cognitive) phenomenon.

At this point, the mechanist might insist that biological realist commitments go hand in hand with the decompositional and constitutive character of mechanism, playing an essential role in distinguishing good from bad mechanistic models of cognitive capacities. In consequence, they should not be omitted from an account of the mechanistic model of cognitive explanation. In line with this intuition, the proposed reconstruction of the moderate realist view of mechanistic explanation articulated in section 3.2 has shown that these sort of biological commitments can be viewed as ontic constraints or norms which, alongside additional epistemic constraints, guide the construction of good mechanistic models of specific features of cognition (cf. Illari 2013). However, given that these constraints are embedded in a network of specific experimental and theoretical practices, they ought not be taken to constitute global norms for all types of approaches devised to study cognitive phenomena. Instead, they should be viewed as 'local' norms that help scientists establish which of the proposed mechanistic decompositions offers an adequate explanation of the phenomena being investigated at a particular level of analysis and/or abstraction.

Recognising that ontological principles play this sort of 'local' normative role in guiding the practice of constructing particular models of cognitive phenomena helps resolve the tension between the

bottom-up character of the decompositional strategy associated with the mechanistic framework and the idea that there are multiple mechanistic decompositions that count as explanatory at different levels of analysis and/or abstraction (i.e., cellular, molecular, biochemical, etc.). Thus, instead of viewing mechanistic ontological commitments as abstract constraints which determine what counts as explanatory for all times and in all contexts, I propose to adopt a strategy which identifies the different ontological principles on which the explanatory practices of different groups of scientists rely. This interpretation not only recovers the most important results of the philosophical analyses of mechanistic models/theories of cognitive phenomena, but also reflects the inherent limitations of applying the mechanistic framework to the study of cognition.

Some mechanists have proposed circumventing these purported limitations by emphasising the integrative potential of the mechanistic framework (Craver 2007b; Craver and Piccinini 2011). They have claimed that mechanism provides the most appropriate template for integrating the cognitive hypotheses and/or theories developed at different levels of analysis or abstraction. That is, as a type of decompositional constitutive analysis, mechanistic models have been taken to cover a variety of levels of analysis or abstraction. The mechanistic decompositions developed at each of these distinct levels of abstraction are in turn said to be constrained by both top-down and bottom-up considerations, and by specific theoretical and experimental principles concerning the spatial, temporal, and active organisation of the components postulated by these different mechanistic models.

Mechanists acknowledge that these componential, spatial, temporal, and active constraints vary from one level of analysis to another. For example, within the different fields which contribute to memory and LTP research, electrophysiologists investigate whether individual synapses are silent or active, focusing on the time-course of electrical activities in nerve cells, biochemists focus on chains of reactions in the cytoplasm, enzyme kinetics and reaction rates, molecular biologists analyse the mechanisms for protein production, while psychologists study rates of learning and forgetting. Each of these fields is taken to possess a series of specialised experimental and theoretical tools which are used in order to develop accounts of specific cognitive patterns/phenomena at different levels of analysis and/or abstraction (cf. Craver 2007b).

However, beyond this methodological autonomy, mechanists argue that it is possible to conceive of the investigative efforts of these different communities of researchers as aiming towards the construction of an integrated multilevel mechanistic theory of cognition. This is supposed to be achieved by the intra- and inter-level integration of the different models and hypotheses developed within the distinct subbranches of cognitive science. What is taken to secure the integration

of these distinct models is precisely the decompositional and constitutive nature of mechanism as well as the fact that models developed at different levels of analysis have to continuously accommodate both top-down and bottom-up constraints. The proposed mechanistic conception is one which promotes multilevel integrated mechanistic explanations of specific cognitive capacities.

Now, the general objection I would like to raise against what seems to be a very attractive picture of the aims pursued in the different fields of cognitive science is that if unification is treated as an index of explanatory power, the mechanistic conception runs the risk of conflating two distinct epistemic virtues of cognitive theories/models. Moreover, by overemphasising the integrationist or unificatory potential of the mechanistic framework, its advocates risk blurring the criteria for determining the explanatory value of particular mechanistic models of cognitive phenomena.

For instance, it is not entirely clear whether, from a mechanistic unificationist perspective, higher-order models of cognitive capacities would also count as proper explanations of cognitive capacities. We have seen that, according to the mechanistic account, only how-actually mechanistic models of cognitive capacities count as genuinely explanatory. But how-actually mechanisms are the ones that satisfy as many ontic and epistemic constraints as possible, and higher-order or abstract models (mechanistic or not) would appear to satisfy fewer constraints than lower-level mechanistic models do. This is primarily why I have claimed that the multilevel model of mechanistic explanation does not settle the question whether abstract (mathematical) models can play a proper explanatory role in the investigation of cognitive phenomena.

This internal tension lies at the heart of the mechanistic position. Although mechanists acknowledge that models/theories developed at different levels of analysis are governed by different epistemic and ontic principles, they also imply that only concrete how-actually mechanistic models are appropriate candidates for cognitive explanations. This contravenes the alleged pluralism entailed by the picture of multilevel mechanistic explanations of cognitive capacities (cf. Craver 2007b; Craver and Piccinini 2011). A genuine pluralist position would attempt to characterise the structure of explanatory models of cognitive capacities in terms that are metaphysically as neutral as possible. Furthermore, an explanatory pluralist position in cognitive science would have to accommodate a wider range of pragmatic and epistemic interests that might guide the development of explanatory models/theories of cognition than the ones which are standardly associated with the mechanistic framework. In consequence, a position that does not attempt to reduce prematurely (i.e., *a priori*) the plurality of the explanatory schemas used at different levels of analysis of cognitive phenomena would have to count the mechanistic strategy

as merely one of the potential schemas that can be used to develop explanatory models of particular aspects of cognitive phenomena.

3.5.2 *Lessons from mechanism*

I conclude by summarising the main lessons that follow from this critical analysis of the mechanistic model of cognitive explanation. First, I have shown that the unrestricted extension of mechanism as the single explanatory category appropriate for the cognitive domain faces a series of difficult challenges. Whilst mechanists standardly appeal to the unificatory or integrative power of the mechanistic framework to sidestep some of these issues, I have argued that this strategy ultimately generates even more indeterminacy with respect to the criteria for evaluating the explanatory value of particular mechanistic models/theories of cognitive phenomena. Instead, mechanistic explanations are better viewed as *one* type of cognitive explanation which is both decompositional and constitutive. Mechanistic explanations reveal in a stepwise fashion certain fundamental features of the cognitive phenomena being investigated and are constrained by a series of specific (local) ontic and epistemic principles or norms.

Second, the arguments developed in this chapter show that there are good reasons to resist adopting a strong realist (ontic) perspective on the problem of cognitive explanation. Abstract ontological commitments to 'real' mechanisms that underlie observable cognitive phenomena generate a number of hard metaphysical puzzles. In addition, this sort of abstract metaphysical commitment to mechanism seems to be an ineffective (conceptual) tool for assessing the explanatory value of current scientific models and/or theories. In other words, the strong ontic constraint seems to be too far removed from the scientific practice to be relevant for the evaluation of particular models and theories of specific cognitive capacities or processes. Instead, the proposed moderate realist view of mechanistic explanation is consistent with the idea that the pursuit of mechanistic models of cognitive capacities is methodologically and epistemically possible and welcome, even in the absence of a mechanistic metaphysics of the mind.

And, finally, conceiving of mechanistic explanation as a complex epistemic activity, that is constrained by both ontic and epistemic considerations, entails that mechanistic explanations of cognitive phenomena are essentially partial. In other words, if constructing good mechanistic models of particular cognitive phenomena depends on the ontological and epistemic constraints one endorses in a particular investigative practice, there is no general way of establishing that all features of cognition will be explainable via the same mechanistic strategy. That is, from a moderate realist standpoint, there might well be aspects of the cognitive phenomena being investigated which are

not amenable to a mechanistic analysis at all. The adoption of a more modest mechanistic perspective is consistent with the idea that there are aspects of cognitive phenomena which are appropriately investigated and explained with the help of other types of tools than those made available by the mechanistic framework.

This brings us to a more general observation which motivates the investigations pursued in the following chapters of the thesis. As with any other broad field of scientific investigation, cognitive science aims to give an account of a variety of cognitive phenomena that spans different levels of analysis or abstraction. The various research projects carried out within cognitive science seek to advance our understanding of general features of cognition, such as: the productivity, systematicity, and inferential coherence of language and thought or the relation between visual processing and abstract reasoning tasks, and the role of affective responses for short and long term decision making, as well as of how different parts of the nervous system react to certain types of internal or external stimuli or of the various patterns of nervous activity which can be observed in different regions of the brain, and of other cellular, molecular, and biochemical patterns observed in the functioning of nervous systems. A general concern, then, stemming from this analysis, is whether all of these cognitive phenomena/patterns can be captured and explained via the sort of decompositional and constitutive analysis characteristic of mechanism.

This point relates to another important question discussed in this chapter, concerning the explanatory value of abstract (mathematical) models of cognitive phenomena. I have argued that there are contexts in which a mathematical structure can be said to play a proper explanatory role with respect to a given cognitive phenomenon. Its explanatory value depends on whether the abstract (mathematical) structure/concept is able to characterise certain fundamental features of the phenomenon under investigation. Although this proposal needs to be articulated in more detail, drawing on the discussion from section 4, I claim that the hypothesis being put forward is consistent with the idea that, in other scientific contexts, abstract models are complemented by mechanistic hypotheses, yielding explanatory models which capture other fundamental features of the target phenomenon.

Whilst the explanatory structure of a mechanistic model consists in its constitutive and decompositional features, the justification of the components and activities postulated by an explanatory mechanistic decomposition depends on the ontic and epistemic norms which connect the explanation of a particular cognitive phenomenon to other theoretical and experimental activities relevant for the investigation of that particular phenomenon. This connection has led some mechanists to claim that the explanatory value of mechanistic models of cognitive phenomena derives in part from their biological plausibility.

In response, I have argued that it is a mistake to conflate the two issues. One *prima facie* reason for resisting this conflation is that claims of biological plausibility are relative to the current state of our biological knowledge which might change (or be partly falsified) in time as our understanding of biological phenomena advances. A more robust mechanistic account of cognitive explanation must distinguish between the explanatory structure of mechanistic models and the 'local' norms (ontic and epistemic) that guide the construction of better mechanistic models. Nevertheless, as we will see in the following chapters, the biological plausibility issue tends to resurface in most debates concerning the problem of cognitive explanation.

One prominent account that confronts the two issues sketched above is the computational model of cognitive explanation. Computational models/theories usually target abstract or general properties of cognitive capacities, while claiming to constitute the first step towards a biologically realistic model of cognition. In the following chapters, I will explore several models of the notion of cognitive computational explanation, seeking to elucidate their relationship with the mechanistic thesis. As suggested in the introduction, one important reason for choosing to focus on this particular cluster of computational models of explanation is that they have often been portrayed as providing a fertile middle ground between our folk-psychological and scientific approaches to mental phenomena. Moreover, computational accounts of cognition have also been taken to constitute the paradigm case of the cognitive revolution in the study of the mind and its place in nature. For these reasons, I take this sort of investigation to provide, despite its limitations, a rich enough base for articulating a novel and more adequate philosophical account of the notion of cognitive explanation.

4

CLASSICAL COMPUTATIONAL EXPLANATIONS

4.1 INTRODUCTION

This chapter and the two that follow focus on three different versions of computationalist approaches to cognition. The main objective of these investigations is to elucidate the structure and implications of some of the most prominent computationalist approaches to cognition. In particular, I seek to show whether and when computational models may be said to have an explanatory function in the context of cognitive scientific research. In pursuit of this topic, I revisit a number of debates concerning some foundational issues related to the central thesis of computationalism, namely that psychological/cognitive capacities are explainable by positing internal computations. In addition to important theoretical clarifications, I seek to bring a distinctive practice-based perspective to these debates by analysing a number of computational models currently developed by practising cognitive scientists.

My aim in adopting this strategy is twofold: (i) to show how certain foundational questions concerning the study of the mind relate to the current theorising and experimental practices of cognitive scientists, and (ii) to identify and correct certain philosophical claims about the principles governing computational modelling and explanation in cognitive science. The objective is to determine whether a view of computational explanation can be developed which is both consistent with current scientific practice and also can be integrated into a broader philosophical picture of the nature of cognitive explanation.

4.1.1 *Classical computationalism: an overview*

I begin the investigation of computationalist approaches to cognition by analysing the case for *classical computationalism*. More precisely, the following analysis aims to elucidate the thesis that cognitive phenomena can be explained in terms of internal computations and operations (rules) defined over them. For this purpose, I will count as 'classical' a host of positions that, in the last forty years or so, have been developed under the banner of the computational theory of mind

(henceforth, CTM). Thus, I will consider a variety of arguments and hypotheses developed by a number of different philosophers and cognitive scientists in support of the idea that classical computationalist principles provide adequate explanatory tools for the study of cognition (e.g., Fodor 1975, 1980, 1987; Pylyshyn 1984; Cummins 1989; Egan 1992, 1995, 2010; Gallistel and King 2009; Shagrir 2001). I submit that by adopting this coarse-grained perspective, one is better placed to show how the core tenets of classical computationalism relate to the current experimental and theorising practices encountered within cognitive science.

As a consequence of treating such a variety of philosophical computationalist accounts on equal footing, classical computationalism will be conceived in what follows as a collection of distinct hypotheses concerning the structure and organisation of particular cognitive capacities rather than a uniform and unitary theory of the nature of the mind. However, the key commitment endorsed by virtually all defenders of classical computationalism is that the *explanation* of cognitive capacities requires the postulation of a system of internal structured representations (symbols) and rules (operations) for manipulating and transforming them. In other words, classical computationalists are all committed to what has been called in the literature a *classical (i.e., rules and representation) cognitive architecture* (cf. Newell 1980a; Pylyshyn 1984; Fodor and Pylyshyn 1988). In addition to this hypothesis about the structure of an appropriate cognitive architecture, classical computationalism comes with a characteristic explanatory strategy, whose main features I will briefly sketch below.

Most versions of classical computationalism endorse either explicitly or implicitly a tripartite picture of computationalist explanation. Perhaps the most influential source for this conception of explanatory levels has been David Marr's (1982) pioneering tripartite view of computational explanation. According to Marr (1982), any computational or information processing system (including human cognitive systems such as the visual system or the language system) can be described at different levels of abstraction. The most abstract level of description corresponds, on his schema, to the computational theory which specifies the task of the system: what the inputs and outputs are, and why. The next level down is the algorithm level which describes the specific representations and operations that are used to accomplish the task described by the computational theory. Finally, the implementation level (theory) describes the hardware underlying the proposed algorithm.[1]

1 Although Marr's (1982) choice of terminology is not the most perspicuous for all aims and purposes, given its influence in the methodological and philosophical literature, I will mostly rely it in the course of the following discussion. Other similar classificatory schemes comprise those proposed by: (1) Alan Newell (1980b) who recognises the knowledge, symbolic, and device levels; and (2) Zenon Pylyshyn (1984), who distinguishes between the semantic, symbolic, and realisation (physi-

Computational approaches to the study of language offer a good example. Following Marr himself, theories of grammar are standardly described as computational-level theories. That is, a computational level syntactic theory is viewed as a description of the inputs and outputs (word strings and hierarchical structure) and the relations between them. Psycholinguistic theories in turn are often regarded as algorithms: they specify the representations and operations responsible for accomplishing the translation between word strings and hierarchical structure. And, at the level of implementation one would need to have an account of how these algorithms are instantiated using neural mechanisms. Whilst the Marrian tripartite picture has been criticised on a number of grounds (e.g., Sun 2008; Phillips and Lewis 2013), it nevertheless can be taken to provide a convenient scheme for assessing the main theoretical tenets of classical computationalism and is broadly followed by virtually all classical computationalists.[2]

Another salient feature of the explanatory strategy associated with classical computationalism is that it consists in a top-down decompositional analysis, similar in some respects to functional analysis (Cummins 1989, 2010) and the mechanistic view of explanation (Craver 2007b; Craver and Piccinini 2011). Computational explanations are decompositional in the sense that they involve the decomposition of a complex cognitive capacity into its simpler component cognitive sub-capacities whose computational structure accounts for important features of the initial target capacity. This characterisation of the general explanatory strategy associated with classical computationalism suggests that there are important elements of continuity between the mechanistic, functional, and computational explanatory schemas. However, this *prima facie* continuity does not necessarily entail the priority of any of the existent explanatory strategies over the others. In fact, I will show that all these explanatory strategies (e.g., functional, computational, and mechanistic) contribute in distinguishable ways to the investigation of particular aspects of cognitive phenomena.

The distinctive mark of the explanatory strategy associated with classical computationalism is that it postulates complex (i.e., internally structured) symbols and rules in order to account for certain salient cognitive patterns/phenomena. As will be argued at length in what follows, most debates concerning classical computationalism conflate two distinct issues which arise in connection with the application of classical (symbolic) computational structures to the study of cognitive phenomena, viz.: (i) the computational individuation issue,

cal) levels. I will resort to these alternative classifications only for the purposes of clarifying certain fine-grained distinctions between the versions of classical computationalism analysed in this chapter.

2 I will return to discuss some of the concerns raised in connection with the application the Marrian picture to the current landscape of computational theories/models of cognitive capacities in the last section of the chapter.

and (ii) the cognitive explanation issue. Whereas the first concerns the criteria that determine what makes something a particular type of computational structure or state, the second issue amounts to clarifying when and why certain computational structures can be said to play an explanatory role with respect to particular cognitive problems.

With regard to the individuation issue, proponents of classical computationalism have been divided between defending a semantic view of computational individuation or a formal view of computational individuation. For the purpose of the following arguments, I will use the designation 'semantic' to stand for *externalist* semantics, that is, semantics that relate a state to things other that its formal computational effects within a computational system, including objects and properties in the external world. It should be noted however that the contents assigned to a state by an externalist semantics are not necessarily only concrete objects and properties in the environment. They may also be abstract (or non-existent) entities and properties. Thus, a semantic view of computational individuation holds that the semantic contents of computational structures determine (at least in part) their computational type-identity. A formal or internalist view of computational individuation, on the other hand, claims that only internal structural properties/relations of a computational system contribute to fixing its computational identity. In addition to the two distinct computational individuation hypotheses, both camps of classical computationalism have developed specific accounts about when and why computational structures might be said to explain particular cognitive phenomena/patterns.

In what follows, I will show that by drawing a systematic distinction between the computational *individuation* and *explanation* issues, one is better placed to evaluate the strengths and weaknesses of the two strands of classical computationalism. Following this strategy, I aim to provide a more consistent account of the structure of classical computational explanations of cognitive phenomena.

4.1.2 *Outline of the argument*

The structure of the rest of the chapter comprises three distinct parts. Section 2 focuses on the issue of classical computational individuation. I analyse the main arguments put forward by classical computationalists for the two distinct positions with regard to what counts as an appropriate individuation scheme for computational systems. The aims of this analysis are twofold: (i) to demonstrate that the semantic (externalist) view of computational individuation oscillates between two incompatible criteria, viz. essential semantic contents and the formality constraint, and (ii) to argue that once the semantic individuation strategy is exposed as being deeply problematic, its

proponents are forced into accepting a non-semantic formal picture of computational individuation, as proposed by the internalist.

In section 3, I turn to investigating the issue of the explanatory value of classical computational models of cognition. I seek to clarify the idea that semantic interpretations (i.e., assignments of mental contents to the internal states of a computational system) play an important 'bridging' role in securing the applicability of abstract computational models to the study of specific cognitive capacities/phenomena. In light of these considerations, I emphasise the advantages of adopting a broadly externalist view of the applicability of computational models to cognition, whilst retaining an internalist stance on the issue of computational individuation.

The last section concludes with an evaluation of the main tenets of classical computationalism. More specifically, I focus on the strengths of the emerging picture of classical computational explanations and suggest a general strategy to circumvent one traditional metaphysical challenge concerning the abstract character of classical computational models/theories of cognitive capacities.

4.2 THE PUZZLE OF COMPUTATIONAL INDIVIDUATION

This section focuses on the issue of computational individuation. As suggested in the introduction, I claim that by keeping the individuation and explanation issues apart, one is in a better position to construct a more robust account of the applicability of computational models to the study of cognition. I aim to show that the principles governing the type-individuation of the internal states and structures of a computational system are distinct (separable) from the principles that determine when and whether a given computational structure adequately describes and perhaps explains a target cognitive capacity. By clarifying the nature of the computational individuation issue, I seek to shed further light on the complex relationship that holds between computational individuation and explanation in the specific context of computational cognitive science and neuroscience.

The puzzle of computational individuation consists in establishing the factors that play a role in fixing the computational identity of a given state or structure. In other words, solving the individuation puzzle requires that one specify the features of a system in virtue of which it belongs to a particular computational type rather than another. Thus, the underlying assumption of the individuation puzzle is that only certain features of a complex system (concrete or abstract) are pertinent to its computational identity. These features (and only these) determine whether two or more structures are computationally indistinguishable or not. If two structures share all their computationally relevant properties, then they are identical from a computational

point of view; otherwise, they are computationally distinguishable, i.e., they count as different computational states or structures.

The philosophical literature on classical computationalism comprises two main proposals for an appropriate computational individuation strategy: (i) the semantic (externalist) and (ii) the internalist view of computational individuation. Whilst the former claims that mental contents (whatever they are) play an essential individuative role, the later promotes a purely formal (syntactic) computational individuation schema.[3] In what follows, I seek to clarify each of these proposals and disentangle some of the confusions concerning the relationship between the notions of computational *individuation*, *identification*, and *explanation*. I begin by surveying the semantic view of computational individuation and some of the major challenges that have been raised against it. Then I analyse the main arguments for endorsing an internalist view of computational individuation. Finally, I assess the theoretical (conceptual) arguments put forward in support of such an internalist perspective against three representative computationalist models proposed in the field of vision studies.

4.2.1 *The semantic view of computational individuation*

The semantic view of computational individuation claims that computing systems and their states are type-individuated in terms of their semantic contents. Thus, if two purportedly computational states/systems have different semantic contents, then they also count as being distinct *qua* computational states/systems. As a result, since differences in semantic contents are taken to generate differences in computational types, computational states and structures are said to possess their contents *essentially*. There are several important lines of reasoning which are taken to support the idea that semantic contents play a crucial role in computational individuation. In what follows, I discuss each of them in turn.

The semantic view of computational individuation seems to receive some *prima facie* support from the standard construal of the notion of computation (cf., Fodor 1975; Pylyshyn 1984). On this standard picture, a physical system S is said to compute a function F only if, under appropriately circumscribed circumstances, the system will always go from one physical state s_i to another physical state s_f such that for every pair of such states $\langle s_i, s_f \rangle$, it is possible to specify a representation function f in such a way that the value associated with the final state $f(s_f)$ is a function F of the value associated with the ini-

3 Because of its insistence that computational individuation is a purely formal (or syntactic) affair, the internalist position has also been taken to illustrate a reductionist stance in the sense that it denies (reduces) the intentional character of computational theories of cognition. There are good reasons to think that this denomination is unhelpful and rather misleading for understanding the implications of the internalist account of computation and this why I will continue to avoid using it.

tial state $f(s_i)$. In changing its internal state from s_i to s_f, the system is said to compute F, in virtue of the fact that the system's behaviour is interpretable under the function f as returning the value of the function F whenever presented with an argument for that function.

Whilst the standard construal of the notion of computation does indeed imply that computations are defined over symbols (i.e., structures that can be assigned semantic contents), it fails to entail that these symbolic structures possess their contents *essentially*, i.e., that contents uniquely determine the computational-type of these structures. In fact, the standard construal is consistent with the hypothesis that the same computational structures can be assigned multiple (i.e., non-unique) semantic interpretations. Put differently, the standard notion of computation affords a clear-cut distinction between symbols and their (semantic) contents. The point here is that the principled separability of the two notions ultimately undercuts the strong support that the standard construal of computation allegedly offers to a semantic view of computational individuation.

The same sort of response strategy can be used to reject a similar line of reasoning that has been dubbed 'the argument from the identity of computed functions' (e.g., Shagrir 1997, 1999; Peacocke 1999; cf. Piccinini 2008a). This argument relies on two main premises: (i) computing systems are individuated by the functions they compute, and (ii) functions are individuated semantically, i.e., by the ordered couples: ⟨domain element, range element⟩, denoted by the inputs and outputs of computation. From (i) and (ii), it is derived that computing systems and their states are individuated semantically. However, this sort of argument is a *non sequitur* because computational functions can be individuated in purely formal or syntactic terms, i.e., non-semantically. The semantic scheme can be used to *identify*, with respect to a particular context, the internal states of a computational system that is taken to perform a particular task. However, given that the semantic interpretation adds extra assumptions about the states of a computing system (e.g., referential assumptions) than those required by the individuation task itself, it is best characterised as an extrinsic *identification* strategy. Thus, as before, the semantic layer is not necessary for the purposes of computational individuation proper.

However, the dismissal of these two closely related lines of reasoning need not pose an insurmountable problem for the defender of a semantic view of computational individuation since she can resort to other, allegedly more pertinent and convincing, argumentative strategies. I what follows, I will consider a set of arguments which may be conveniently grouped together in virtue of the fact that they explicitly equate the computational individuation and explanation issues. The common strategy underlying these arguments starts from the observation that internal computations are postulated in order to explain various interesting cognitive phenomena. The next step is to infer

that at least some of the properties that play a role in determining the explanatory value of computational models of cognitive phenomena also contribute to fixing their computational type-identity.

One of the most intuitive versions of this type of argument rests on the assumption that computationalist theories of psychological capacities must be able to vindicate (at least in part) the explanatory success of folk-psychology. The latter in turn is taken to depend to a large extent on the fact that folk-psychology individuates psychological states such as beliefs, desires, and thoughts in terms of their contents and thus succeeds in capturing a large number of interesting psychological patterns/behaviours. For instance, the commonsense (folk-psychological) explanation of my tea-drinking behaviour appeals to my desire to drink tea (rather than coffee), my belief that there are tea-making ingredients nearby (rather than at the supermarket), and my ability to prepare tea. In line with this, it has been proposed that, if computational psychology is to vindicate this sort of psychological pattern, then it must also individuate its theoretical computational posits in terms of intentional (semantic) categories. Despite its *prima facie* intuitive appeal, almost everyone agrees now that whatever (complex) relation holds between commonsense psychological explanations and computationalist explanations of cognition, it is not one that forces computational psychology to import the categories and tools used by folk-psychology to generate certain patterns of explanation. Otherwise put, the expectation that folk-psychological taxonomies will simply be imported into computationalist psychology is not likely to survive even brief scrutiny of past and current scientific practice in and outside of the arena of cognitive science.

Still, one need not rely on the fact that folk-psychology individuates psychological states intentionally, i.e., in semantic terms, for there is a much more compelling case to be made that large parts of scientific psychology and cognitive science itself make use of such intentional (semantic) vocabularies on a daily basis. This seems to bolster the claim that mental states and processes are indeed individuated in semantic terms, which in turn leads to a new argument for semantic computational individuation. Assuming both that: (i) the identity of mental states (capacities, processes, etc.) depends essentially on their semantic properties, and (ii) computationalist theories of cognition propose to explain precisely these psychological capacities, it follows that the theoretical posits of computational theories of cognition must also be individuated in semantic terms (cf. Fodor 1975; Burge 1986; Peacocke 1999; Wilson 1994). The main assumption driving this style of argument is that, in the case of scientific explanations of cognitive phenomena, the *explananda* and *explanantia* must be individuated in similar, i.e., semantic, terms. But, as pointed out above, there seems to be a straightforward problem with this pro-

posal. Whilst it is true that one must usually be able to provide an account of how a particular theory relates or applies to a given set of observable phenomena (or to the pre-theoretically specified questions raised in a particular domain of inquiry), there is no reason to expect that such an account will take the form of a one-to-one mapping between the 'pre-theoretical' categories and the posits of a particular scientific model/theory.

This conclusion can be reinforced by invoking a more general argument that has been put forward against the idea that contents play a crucial role in computational individuation. In what follows, I analyse two slightly different versions of this argument which is originally credited to Stephen Stich (1983, 1991). The first step in Stich's general *eliminativist* argument[4] against the individuative role of mental contents is to point out that content ascriptions have a series of properties (i.e., *R* (for relativity)-properties, cf. Egan 2009) which make them inadequate tools for computational individuation. More specifically, content ascriptions are said to be both vague and context-sensitive. That is, for any given predicate of the form 'believes that *p*' there will be some contexts in which it applies and others in which it does not apply, thus making it indeterminate whether a generalisation that invokes such a predicate will hold or not across all contexts and conditions. Appeals to content are also observer-relative in the sense that: 'to believe that *p* is to be in a belief state similar to the one underlying our own sincere assertion of *p*' (Stich 1983, p. 136). In addition, appeals to content often seem to presuppose ideological similarity as well as reference similarity. Two beliefs are ideologically similar if and only if they are embedded in similar doxastic networks, and they are reference similar only to the extent that the terms subjects use to express the beliefs have the same referent.[5] Because content ascriptions have the *R*-properties, Stich argues that they impose a more fine-grained individuative scheme than seems appropriate for use in any field of scientific psychology, which in turn is taken to imply that they are not adequate tools for computational individuation.

Stich's (1983) original argument against the semantic view of computational individuation was formulated in terms of a specific constraint which he dubbed the *Autonomy Principle* (AP). According to AP, 'any state or property invoked in a psychological explanation

4 Traditionally, Stich's (1983, 1991) anti-content arguments have been read as implying a strong eliminativist position, according to which mental contents do not play any role in computational models of the mind. I think there are good reasons to reject this eliminativist reading and I provide further support for this claim in the following sections. For present purposes, it suffices to point out that a strong (eliminativist) reading of Stich's anti-content conclusion risks to conflate the individuation and explanation issues, against the policy stated in the introduction.

5 Thus, Jo's belief that 'Huw is a conservative' and Don's belief that 'Huw is a conservative' count as the same belief if and only if all of Jo's relevant beliefs about Huw and conservatives are the same as those of Don; and the same goes for their referential intentions.

should supervene on the current, internal, physical state of the organism' (Stich 1983, p. 164). Taken at face value, AP provides a good illustration of the major motivations driving the classical debate in philosophy of mind concerning the distinction between *broad* (wide) and *narrow* mental contents and their respective roles in a computational theory of mind. Unlike broad contents which are taken to depend on the subject's historical, environmental or social context, narrow contents are standardly taken to be supervenient on the physical states of a system.[6]

In brief, AP claims that the sort of scientific explanations of cognitive behaviours sought by cognitive scientists and psychologists should apply to all physical duplicates of an organism. That is, such explanations should invoke only narrow states and properties (shared by all physical duplicates); in particular, Stich claimed that scientific cognitive explanations should invoke only narrow contents. Thus, this line of argument leaves open the possibility that some notion of narrow mental content impact computational taxonomies. Otherwise put, Stich's (1983) *anti-content* argument seems to be consistent, *pace* AP, with the hypothesis that narrow contents play (in part) an individuative role in computational psychology.

But while Stich's original argument did focus on the notion of ordinary (externalist) content (i.e., the sort of content ascribed in folk-psychological predictions and explanations of behaviour), the revised argument from 1991 claimed that narrow content - which abstracts away from the subject's historical, environmental, and social context - is nonetheless still too vague and context-sensitive (i.e., has too many of the *R*-properties) to determine the taxonomy of computational models of cognitive capacities.[7] The latter, Stich argued, individuate computational states/structures in terms of their *narrow causal role*, which, he insisted, is different from narrow content proper. That is, narrow causal roles are taken to depend only on the internal organisation of the component parts of a computational system which performs a specific function.

There are two preliminary conclusions that follow from Stich's anti-content arguments. Firstly, neither a broad nor a narrow notion of

6 There are several conceptions of narrow content available on the philosophical market. Some authors conceive the narrow content of a mental state such as a belief as the *detailed description* of that particular belief (cf. Putnam 1975; Mendola 2008). A different approach identifies narrow contents with *conceptual roles* (cf. Block 1986). Other influential conceptions of narrow content include the *mapping account* (Fodor 1987), Stalnaker's notion of *diagonal propositions* (Stalnaker 1990, 1999), and Chalmers' notion of *sets of maximal epistemic possibilities* (cf. Chalmers 1996, 2002).

7 To summarise, the argument against narrow content says that: 'The categories of a narrow content taxonomy are simply the categories of a broad content taxonomy extended to meet the demands of the principle of autonomy. But the broad content taxonomy of commonsense psychology is too vague, too context-sensitive and too unstable to use in a serious scientific theory. *Narrow* content inherits all the deficits' (Stich 1991, p. 250).

mental content seems to be well-suited to play an individuative role in computational theories of the mind. Secondly, despite their obvious differences, the two types of content have many features in common (i.e., their common *R*-properties), which implies that whatever roles they might turn out to play in computational theories of cognition, they will, in any case, be very similar. In the remainder of this section, I survey two more strategies that have been used to promote a semantic view of computational individuation.

Fodor (1975, 1987) and Pylyshyn (1984) have argued that one of the main attractions of CTM is that it seems to provide a straightforward account of the 'striking parallelism' between the causal relations among propositional attitudes and the semantic relations connecting their particular contents. The key ingredient of the purported success of CTM is the postulation of internal representations that have a dual character, i.e., they have formal properties which are taken to be causally efficient (in virtue of the fact that they represent classes of physically equivalent states) and can be assigned semantic interpretations (contents). Furthermore, according to these authors, in order to explain such a striking parallelism one is led to suppose not only that causal powers are attributed to internal states that are taken to be semantically evaluable, but also that 'causal relations among propositional attitudes somehow typically contrive to respect their relations of content' (Fodor 1987, p. 12).

If the latter assumption is correct, then it seems that contents must play at least a *partial* role in the type-individuation of these states. For if mental contents do not play any role in the type-individuation of computational states, then states with different semantic contents would still count as being the same from a computational point of view. But then, arguably, the causal relations holding between computational states would not mirror perfectly the semantic relations holding between their contents, contrary to the strong parallelism view. A similar line of reasoning can be found in Pylyshyn (1984). Although he admits that in general computational states can have multiple semantic interpretations, he maintains that '[i]n the case of a psychological model we want to claim that the symbols represent one content and no other, since *their particular content is part of the explanation of its regularities*' Pylyshyn (ibid., 40, m.e.). Therefore, the main conclusion of this type of argument seems to be that representational structures posited by a computational theory of *cognition* must have essential (unique) contents which participate (partly) in their type-individuation *qua* computational structures/states.

There are three interrelated points to raise in response to this line of argument. First, the conclusion being put forward relies on simply equating the factors that play a role in determining the explanatory value of particular computational models/theories of cognitive capacities with the factors that fix their computational type-identity. This

sort of conflation is problematic, especially in light of the fact that the authors recognise that in different contexts (i.e., outside cognitive modelling), the individuation of computational systems/structures does not depend on the semantic contents assigned to them (cf. Fodor 1980; Pylyshyn 1984). Given that an adequate view of computational individuation should presumably provide a uniform account of what makes something a computational state/structure that is independent of the specific contexts in which computational structures are being used and/or applied to study various types of problems, there are good reasons to resist the adoption of the semantic individuation strategy proposed by these authors.

Second, granting that the 'syntax parallels semantics view' captures a series of salient empirical features of cognitive phenomena which require an adequate explanation, one still needs to establish the extent to which this parallelism holds in the cognitive domain, i.e., whether it holds across all types of cognitive phenomena or only for a limited range of them. Otherwise one could be criticised for trying to establish what should count as the correct view of computational individuation based on a series of sweeping speculative considerations concerning the nature and structure of this purported parallelism. Both of these points are intended to highlight the limitations of extracting a view of computational individuation solely from the modelling practices of computationalist cognitive scientists.[8]

Finally, the argument strategy that seeks to secure a semantic view of computational individuation based on the explanatory interests of practising cognitive scientists seems to be in conflict with another important criterion for computational individuation discussed in the philosophical literature, namely the *formality constraint* (cf. Fodor 1980). In a nutshell, the formality constraint requires that the internal states/structures and operations postulated in a computational model of a particular cognitive capacity have formal properties which are mapped via an adequate function in equivalence classes of physical properties, which in turn explains why these properties are taken to be the ones that are actually effective in the mechanisms underpinning cognitive processes.

Thus, according to the formality constraint, the representations and operations postulated by a computational model/theory of a particular cognitive capacity are individuated in terms of a set of structural (formal) relations which hold between the component parts of the system. What counts for the individuation of a system/structure *qua* a certain type of computational system are neither the particular (external) contents that can be assigned to its states nor the specific physical properties of the system that might implement that particular type of

8 As will be argued in the following sections, a more adequate account of the factors that determine the computational type of a particular state/structure should be relatively orthogonal to the particular epistemic aims (e.g., explanation, prediction, confirmation) of practising cognitive scientists.

computation. Instead, from a computational point of view, what distinguishes between different types of computational structures are their structural (formal) properties.[9] As pointed out by Fodor (1980) himself, the formality constraint is quite strict because it implies that mental contents are not *sufficient* for individuating the internal states of a computational mechanism:

> [I]f the *computational* theory of the mind is true (and if, as we may assume, content is a semantic notion par excellence) it follows that content alone cannot distinguish thoughts. More exactly, the computational theory of mind requires that two thoughts can be distinct in content only if they can be identified with relations to formally distinct representations. More generally: fix the subject and the relation, and then mental states can be (type) distinct only if the representations which constitute their objects are formally distinct (Fodor 1980, p. 64).

However, it might be argued that this formulation of the formality constraint is still compatible with the idea that contents play a *partial* role in the individuation of computational states. As an example of a sophisticated argument put forward in support of this sort of claim I will briefly consider the position defended by Oron Shagrir (2001). As per the previous arguments, Shagrir discusses the problem of computational individuation in the context of understanding computationalist approaches to cognition. He claims that certain computational systems (models) that satisfy one formal (syntactic) description will generally satisfy other formal descriptions as well (cf. *the multiplicity of syntactic implementation* assumption). Shagrir (2001) further argues that, when considered outside a specific modelling context, one cannot decide which of the possible formal descriptions of a computational model is appropriate with respect to the cognitive task being modelled. He then concludes that in order to settle which formal description individuates the computational system of interest one needs to take into account the semantic description of the task being modelled. This in turn is taken to imply that semantic interpretations (or contents) play a partial role in the individuation of computational models of cognition, i.e., they help select the relevant individuative formal description.

But upon closer scrutiny it turns out that this argument does not support a semantic view of computational individuation after all. For in specifying the semantic features which *do* play a role in computational individuation, Shagrir (ibid., p. 20) writes that: 'I do not claim

9 Some authors (e.g., Piccinini 2007) have argued that the formality constraint should not count as a proper individuative principle because it is formulated in terms of the causal powers of computing mechanisms. However, I believe that the reference to causal powers does not affect the idea that only structural or formal properties of a computational system play an individuative role.

that *every* change in content alters computational identity. The features that make a computational difference, in my view, are *formal features*, that is, set-theoretic relations and other high-level mathematical relations among the represented objects.' (m.e.) Although such formal features might be taken to constitute a kind of *internal* semantics of the computational system, they do not seem to relate the states of the system with something outside the system, thereby qualifying as semantic contents in the standard externalist sense. Rather, formal features can be seen as characterising internal relations between the components of a computational system. This brings Shagrir's (2001) account closer to the internalist view of computational individuation which will be analysed in more detail in the following section. Nevertheless, one might insist that the *identification* of the right formal features that play a role in the type-individuation of a particular computational system is facilitated or mediated by the externalist semantic description of the task performed by the system. Although this latter point may turn out to be valid, it does not entail that semantic (externalist) contents play a role in the computational individuation of a particular system or structure.

Finally, a somewhat different line of reasoning aims to establish that (mental) contents play a role in the individuation of computational systems and their states by showing that prominent computational models of cognitive capacities adopt (explicitly or implicitly) precisely such a semantic individuative strategy. Tyler Burge (1979, 1986, 2010) is one of the authors who has most forcefully pursued this defence strategy. He claims that the modelling practices of computationalist psychologists in fields such as vision studies support a semantic view of computational individuation. Whilst Burge's analyses reveal a number of important aspects of the complex epistemic activities of constructing adequate explanatory accounts of particular cognitive capacities, I contend that such a practice-based perspective does not by itself entail the semantic view of computational individuation. Instead, I argue that some of Burge's most interesting criticisms of the internalist view of computationalism, which rely on the analysis of the modelling practices of vision scientists, pertain more directly to the explanatory value of computational models/theories of cognitive capacities. I will return to the analysis of these insights and their implications, in section 2.3, after evaluating the main theoretical considerations that have been proposed in support of an internalist view of computational individuation.

4.2.2 *The internalist view of computational individuation*

As a version of classical computationalism, the internalist account of computation endorses the hypothesis that computational models of cognition postulate internally structured representations (symbols)

and operations (rules) defined over them. However, defenders of the internalist account also insist that the interpretability of the mental symbols postulated by classical computationalist models does not entail that these symbols possess their (semantic) contents essentially. That is, the internalist holds that, although postulated computational states may be assigned specific/particular contents, these are not essential for their *computational identity*.

In what follows, I will use the account that Frances Egan has developed and refined in a series of papers (cf. Egan 1992, 1999, 2010, 2013) as a starting point for discussing the strengths and weaknesses of the internalist position within classical computationalism. However, as will be shown in what follows, although Egan's position is advertised as an internalist view of computational individuation, there are good reasons to question the overall consistency of her account. In line with the formality constraint (Fodor 1980), Egan holds that an adequate computational individuation strategy should follow only very abstract or formal principles. She spells out this picture of computational individuation by claiming that a computational system (state) is individuated by two types of mappings: (i) a realisation function (f_R) which maps a particular computational process into an appropriate equivalence class of physical states and (ii) an interpretation function (f_I) which maps the computational process into a canonical description of the function performed by the system. The latter mapping, Egan claims, is a mathematical function which together with the realisation function contributes to the *type-individuation* of the computational system.

In contrast to a more traditional semantic account (e.g., Newell 1980a; Pylyshyn 1984) where the interpretation function (f_I) provides a mapping between equivalence classes of physical states of a system and elements of some external (or internal) represented domain, Egan (2010) insists that the interpretation function merely provides a canonical mathematical description of the function computed by a particular type of physical system. She argues that the interpretation function, thus conceived, *does* play an individuative role, and in virtue of this property, it should be distinguished from a proper (externalist) semantic interpretation:

> The characterisation of a computational process or mechanism made available by the interpretation function f_I - the mapping that provides a canonical description of the function computed by the mechanism, and hence (along with the realisation function f_R) serves to *type-individuate* it - is [...] an abstract mathematical description. This semantic interpretation does not provide a *distal* interpretation of the posited internal states and structures; the specified domain is not external objects and properties [...] but rather mathematical objects. The interpretation maps the states

> and structures to a domain of *abstracta*, hence the specified relation is not regarded, in the theory, as a Naturalistic relation. It cannot be a causal relation since abstracta have no causal powers (Egan 2010, p. 256).

Hence, the revisionary twist in Egan's formulation of the individuation criterion consists in modifying the traditional domain of the interpretation function, from objects and features of the external or internal environment of the system to mathematical structures. Taken at face value, the idea that the interpretation function (f_I) offers a mathematical characterisation of the task performed by the system is in line with the notion that a computational theory is meant to give an abstract or highly general account of cognitive phenomena. The mathematical characterisation provides a concise way of *identifying* the parameters which are essential from a computational point of view (thus reducing the degrees of freedom of the system). This idea does not conflict with the notion that a computational process is just an abstract operation, defined over appropriate types of symbols, which is neutral with respect to the actual content of the symbols manipulated, transformed or created in the course of computation.

Egan (ibid.) reinforces her point about the formal individuation of computational mechanisms by appealing to two examples from computational cognitive science. The first example is taken from Marr's (1982) traditional computational theory of early visual processing, in which the computational mechanism for the initial filtering of the retinal image is described from a computational perspective in terms of two specific mathematical functions. More specifically, the device is said to compute the function $\nabla^2 G * I$ (the X channels) and its time derivative ($\frac{\delta}{\delta t}(\nabla^2 G * I)$ (the Y channels). The second example is taken from the computational neurobiology of reaching and pointing. The model of object manipulation proposed by Shadmehr and Wise (2005) decomposes the computational task of object grasping in three subtasks, each of which is amenable to a mathematical canonical characterisation. In relation to these examples, Egan comments that:

> [t]he important point is that in both examples the canonical description of the task executed by the device, the function(s) - computed, is a mathematical description. [...] this description characterises the mechanism as a member of a well-understood class of mathematical devices. A crucial feature of this characterisation is that it is 'environment neutral': the task is characterised in terms that prescind from the environment in which the mechanism is normally deployed (Egan 2010, p. 256).

The last part of the previous quote insists that the proper computational description of a target cognitive system is highly abstract (i.e., mathematical) and that any externalist semantic interpretation

attached to it provides only an extrinsic description of the computational system. As such, it does not affect the type-individuation of the internal states of the computational system, i.e., it does not play any part in the individuation process.

Nevertheless, as will be shown in section 3.3, on Egan's account, the semantic (externalist) description does play several important, albeit non-individuative, functions in the construction and assessment of successful computational models of specific cognitive capacities. For this reason, the formal individuation hypothesis should not be mistaken for the claim that in building a computational theory of cognition, practising cognitive scientists ignore the various contextual (environmental) factors which shape the specific computational problems that cognitive systems are taken to solve. As Egan (1999) herself points out, this sort of blindness to mind-external world interactions would make computational structures inappropriate tools for the modelling and explanation of genuine biological capacities. However, she correctly points out that this observation is consistent with the idea that the computational characterisation of a target cognitive phenomenon is itself 'environment neutral' or formal.

Therefore, I take Egan's arguments for a formal (internalist) view of computational individuation to pursue two interrelated aims: (i) to reinforce and sharpen the main reasons for resisting the adoption of a semantic view of computational individuation and (ii) to promote the idea that computational models/theories provide an abstract or very general description of certain target features of particular cognitive phenomena. Besides the theoretical considerations she brings in support of her position, Egan claims that the modelling and theorising activities of computationalist cognitive scientists support an internalist view of computational individuation. There are two main concerns that I would like to raise in relation to Egan's internalist position.

Firstly, a number of authors have argued that Egan's view of computational individuation does not actually qualify as a proper non-semantic account because mathematical contents, which, on her view, play an individuative role, constitute a thin semantic layer after all (cf. Shagrir 2001; Piccinini 2007b). That is, from a purely formal perspective, the different kinds of contents assigned by an interpretation function are all equivalent (i.e., each of them constitutes a range or co-domain for the interpretation function). As such, Egan's account seems to be very close to Shagrir's (2001) semantic view of computational individuation. For, as we have seen above, Shagrir claims that only formal features ('set-theoretic or other higher-order mathematical relations') determine the computational type of a particular computational system/state. So, it seems that either both views should be qualified as being semantic accounts or they should both count as internalist accounts of computational individuation.

Secondly, as mentioned above, Egan (1992, 1995, 2010) has argued, contra Burge (1979, 1986) and others (e.g., Davies 1991, Silverberg 2006), that the scientific practice of constructing good explanatory computationalist models of cognitive capacities supports an internalist view of computational individuation. As a response, I contend that the modelling practices of computationalist cognitive scientists cannot serve as an arbiter in disputes concerning the correct view of computational individuation. This is because, as will be shown below, most of the principles that guide the practice of constructing good explanatory models of cognitive capacities do not bear directly on the computational individuation issue. However, this does not invalidate the idea that the correct view of computational individuation is one that mandates that the computational type-identity of a particular system depend exclusively on the formal (structural) properties and/or relations of its internal component states.

Thus, I claim that Egan's account fails to provide proper internalist (formal) criteria for computational individuation. As stated above, the two factors that generate the tension which lies at the heart of Egan's (2010, 2013) account are: (i) the thesis that mathematical (interpretation) functions play a role in the type-individuation of computational systems, and (ii) her strong reliance on the modelling practices of computationalist psychologists. Whilst in section 2.3 below, I will show some of the main difficulties of taking computational modelling practices in the cognitive domain to guarantee an internalist view of computational individuation, in section 3.2, I will attempt to provide a more adequate treatment of the role(s) played by mathematical contents/structures in constructing adequate computational models of cognitive phenomena.

In summary, the theoretical considerations discussed so far entail an internalist view of computational individuation according to which the type-identity of a computational system depends only on certain formal (structural) properties and/or relations of the internal components of the system. On a more traditional formulation, the computational identity of a system is determined solely by the realisation function which maps classes of physically equivalent states or properties into certain formal (structural) properties or relations. Having established that the internalist or formal view is the correct way of thinking about the issue of computational individuation, there is a more difficult issue which needs to be addressed next. This issue concerns the factors and/or principles which play a role in determining whether and when a particular computational structure can be used to model and explain a particular cognitive phenomenon.

In what follows, I propose to describe briefly three models from vision studies, in order to identify some of the lessons that the current scientific practice of computationalist modelling affords for a more general philosophical analysis of classical computationalism. In par-

ticular, I claim that such a practice-based perspective affords important insights about the roles that both mathematisation and an externalist semantics play in the construction of good classical computational models of cognitive phenomena. For this reason, the following section also constitutes a good entry point for the discussion of the issue of the *explanatory value* of computational models of cognitive phenomena.

4.2.3 *Computational modelling in practice*

I will focus on three examples of computational models from vision studies: (i) a model of early visual processing that suggests why neurones in the initial stages of the visual pathway have the particular filter properties that they do; (ii) a model that links early vision with object recognition; and (iii) a model of how object recognition might influence early vision. Although much more complex and accurate computational models are available in the literature, the three models discussed here serve a useful expository purpose due to their simplicity. For instance, they illustrate a number of strategies used by practicing scientists to overcome some of the most important challenges of the computational modelling of vision, e.g., the poor input quality, the severe underconstrainedness of the problems, the necessity of rapid computation of results, and the need for incorporating high-level or cognitive influences in the computations. Moreover, considered together, these three models cover a broader range of visual processes, thus providing a more general perspective on what is involved in the computational modelling of a specific domain of cognition.

Visual processing can be partitioned broadly into two stages - an early stage concerned with image representation in terms of a basic vocabulary of filters, and a 'late' stage concerned with recognition. To understand the connection between early visual areas and downstream recognition processes, the analysis must begin with the current understanding of what is being computed in early vision and why. A wide range of computational and physiological studies on early vision have converged towards characterising the function of this part of the visual system in terms of edge extraction. Descriptive models of early visual receptive fields usually take the form of Gabor patches or wavelets, both of which provide a means of representing image structure in terms of local oriented edges at multiple scales. In addition, the selection of this particular representational scheme is supported by the fact that, across a wide number of studies, wavelet-like structures emerge as a robust solution to visual redundancy reduction (cf. Bell and Sejnowski 1997; Olshausen and Field 1996, 2005). It has been claimed that the wavelet-like structures are a plausible representational format for the computations performed by

V1 cells because they are compatible both with the hypothesis that any sensory system constructs representations that take advantage of the latent structures in the input to obtain an efficient code for incoming stimuli as well as with the fact that natural scenes processed by the visual system are highly structured. Despite an ongoing debate over how well simple Gabor or wavelet models of V1 receptive fields describe visual processing, the modelling community has endorsed such features as a useful preliminary step in a wide range of computational tasks, amongst which one can also count object recognition. More precisely, for object and face recognition, representing images with a multiscale 'pyramid' of oriented edge information or information closely related to this has become a standard pre-processing step for many successful algorithms.

The 'qualitative' computational model of face recognition (cf. Sinha 2002; Sinha and Balas 2008), the second of the three models under consideration, starts from the observation that neurones in the early stages of the visual pathway are sensitive primarily to ordinal, rather than metric, relations.[10] Thus, in the representational scheme proposed by the model, objects are encoded as sets of ordinal relations across large image regions. The model aims to explain how ordinal encoding can permit face recognition despite significant appearance variations. It does so by identifying a series of local stable ordinal measurements that encode stable facial attributes across different illumination conditions. By combining all these invariances, a larger composite invariant is obtained, called a ratio-template in virtue of the fact that it comprises a set of binarised ratios of image luminance.

The computational problem addressed next by the model is how the structure of the ratio-template (which constitutes the representation of a face under different illumination setups) is matched against a given image fragment to determine whether it is a face or not. The model postulates two computational stages: (i) averaging the image intensities over the regions laid down in the ratio-template's design and determining the prescribed pair-wise ratios; (ii) determining whether the ratios measured in the image match the corresponding ones in the ratio-template. The latter problem is basically treated by the modellers as an instance of the general graph problem. In brief, the model assumes that the visual system solves this sort of computational problem and determines an overall match metric which is used to establish whether the given fragment image contains an image or not.

This 'qualitative' model of face recognition further implies that the computational processes underlying visual detection tasks do not comprise the extraction of 3-D shape and other related features of the perceived image, but rather rely directly on outputs of the low-

10 Because of this dependence on ordinal relations, the units of the model are called qualitative in their responses.

level visual mechanisms.[11] However, this leaves open the problem of the difference between the high-quality of our visual percepts and the indeterminacy of the low-level visual outputs. The last model to be sketched here aims to provide an account of how edge extraction by the early visual system can yield stable visual percepts under a wide range of varying conditions. Its main assumption is that this computational task is possible only if there are top-down recognition influences that drive the early stages of visual processing like edge-detection, implicit and explicit 3-D shape recovery, colour constancy, and motion analysis (cf. Jones et al. 1997).

Jones et al.'s (1997) computational strategy for incorporating high-level influences in perception uses the concept of *flexible models*. 'A flexible model is the affine closure of the linear space spanned by the shape and the texture vectors associated with a set of prototypical images.' (cf. Sinha and Balas 2008, p. 627) An optical flow algorithm is then used to obtain pixelwise correspondences between a reference image and the other prototype images. These correspondences then serve to represent an image as a 'shape vector' and a 'texture vector'. 'The shape vector specifies how the 2-D shape of the example differs from a reference image and corresponds to the flow field between the two images. Analogously, the texture vector specifies how the texture differs from the reference texture.' (cf. ibid., p. 627) The linear combination of the example shape and texture vectors constitutes the flexible model for an object class. By optimising the linear coefficients of the shape and texture components one obtains the matching of the model to a novel image. The parameters of the flexible model which have been estimated in this way can be used for effectively learning a simple visual task, like 3-D shape recovery and other supposedly early perceptual tasks, such as edge-detection, colour constancy, and motion analysis. Given its general features, the computational model proposed by Jones et al. (1997) is better viewed as an example of a class of algorithms that can be used to learn visual tasks in a top-down manner, specific to different classes of objects.

The three models briefly described above afford two indirect lessons for the problem of computational individuation. The first model of early visual processing illustrates the fact that developing adequate computational models of specific aspects of cognitive processing involves various heuristic and theoretical principles. These also mediate the potential semantic interpretations attributed to different parts

11 Experimental tests have shown that, despite the various idealising assumptions included in the model, the probability of false positives is actually quite small. In addition, the computations postulated by the model are sufficiently straightforward to be executed rapidly. That is, the model does not raise a *prima facie* implementational or realisation problem. This is primarily due to the fact that, against the standard Marrian conception, this qualitative model of face recognition does not require extensive pre-processing of the input image. It directly makes use of the outputs of 'early' visual features, dispensing with the need for complex and error prone computations, such as 3-D shape recovery.

of the computational model (e.g., interpreting wavelet-like features as 'edges'). Otherwise put, mental content ascriptions on this model are indirect and thus cannot be said to play any proper individuative role with respect to the component states and structures of the computational model. The lesson of the first model, therefore, seems to be that semantic interpretations are extrinsic to the computational model *per se* and as such cannot impact the individuation of computational systems.

Taken at face value, the second 'qualitative' model of object recognition might seem to be more supportive of the semantic view of computational individuation, because there are more features of the model being correlated with distal (externalist) features of the perceived objects (e.g., faces). However, a closer scrutiny of the model reveals that the contents being assigned to different components of the computational model constitute a rather mixed bag (i.e., alongside externalist features, the model also mentions contents defined in terms of the outputs of other internal computational systems, and even purely formal features of the model itself). Moreover, the computational structures used at different stages in the model are individuated in virtue of their formal properties, which further undercuts their purported support for a semantic individuation scheme.

Finally, the third model from vision studies highlights the fact that the development of computational models of cognition is usually driven by various theoretical considerations that sometimes go beyond local explanatory concerns. One might seek to develop models that are better integrated with existing prominent hypotheses or that are simpler, or models that can be used to describe and explain patterns which govern the functioning of other, apparently disparate, cognitive systems. Since all these aims and concerns are likely to influence the semantic interpretations assigned to specific computational models, it seems that contents are too indeterminate and context-sensitive to impact computational individuation.

In sum, the analysis of these three models from vision studies seems to undermine the hypothesis that content plays an essential role in the computational individuation of the systems/structures used to model certain cognitive phenomena. However, as will be argued below, I do not take this conclusion further to imply that computational theories of vision are purely formal or internalist. Establishing the features that determine the individuation of particular systems *qua* computational systems is only a part of the analysis of classical computationalist approaches to cognition. What the previous analysis of the examples from vision studies has shown is that there are a range of considerations which drive the construction and refinement of good computationalist models of cognitive phenomena.

In line with these observations, I think that it is possible to mitigate the dispute between 'individualists' (e.g., Chomsky 1995; Egan

1992, 1999, 2010) and 'anti-individualists' (e.g., Burge 1979, 1986, 2010, Shagrir 2001; Silverberg 2006; Piccinini 2008a) by acknowledging that whilst the former attempt to articulate a distinct internalist (formal) view of computational individuation, the latter emphasise the various factors which guide the construction of adequate and potentially explanatory computational models of cognitive capacities. In this way it becomes possible to see that the anti-individualist's appeal to the modelling practices of cognitive scientists is more adequate in the sense that these practices, rather than bearing on the issue of computational individuation directly, provide important insights on the sort of elements and/or principles that practising cognitive scientists take into account when constructing explanatory computational models of particular cognitive phenomena. In what follows, I propose to offer a more systematic treatment of these types of considerations in connection with the problem of the *explanatory value* of computationalist models/theories of cognitive phenomena.

4.3 THE PUZZLE OF COMPUTATIONAL EXPLANATION

The previous arguments led to the conclusion that the internalist view of computational individuation cannot by itself guarantee an appropriate comprehensive analysis of the modelling and theorising practices encountered in different branches of cognitive science. In addition, one needs to clarify what makes certain computational systems appropriate for modelling and explaining particular aspects of cognitive phenomena. For this reason, I will focus next on the structure of classical computationalist explanations and elucidate the sort of principles and norms that guide the development of good computationalist explanations of cognitive phenomena. In this context, I discuss in more detail the role played by the notion of *mental content* in the construction and evaluation of particular computationalist models of cognitive phenomena. I point out that both the semantic (anti-individualist) and the internalist (individualist) versions of classical computationalism appeal to the notion of mental (semantic) content specifically in the context of explanation. This supports the hypothesis suggested above that the semantic interpretation of computational models of particular cognitive capacities is governed by a set of epistemic and pragmatic principles that go beyond computational individuation concerns. More importantly, I contend that the internalist computational individuation strategy advocated in the previous section is a necessary condition of an adequate picture of classical computational explanations of cognition.

4.3.1 *Computational explanations on the semantic view*

The previous sections have argued that mental contents do not impact computational individuation *per se*. However, the key intuition driving most versions of classical computationalism is that mental contents play an essential role in developing appropriate and potentially explanatory computational theories/models of cognition. In order to have a better grasp of the role(s) that mental contents play in a computational theory of cognition it is helpful to consider some of the main motivations that have been put forward for postulating them in the first place. In the specific case of the semantic view of classical computationalism, mental contents are typically introduced in relation to the problem of cognitive explanation. For instance, some proponents of the traditional semantic view of classical computationalism have explicitly argued that a proper level of semantic analysis that postulates mental contents is required in order to be able 'to state generalisations concerning the behaviour of systems under certain descriptions'. That is, because of its power to capture/express certain interesting cognitive patterns, the level of analysis which postulates mental contents is taken to be different from other levels of analysis:

> [I]n a cognitive theory, the reason we need to postulate representational contents for functional states is to explain the existence of certain distinctions, constraints, and regularities in the behaviour of at least human cognitive systems, which, in turn, appear to be expressible only in terms of the semantic content of the functional states of these systems (Pylyshyn 1984, p. 38).

Put another way, for the semantic level to constitute an independent level of description and explanation in a computational theory of cognition, one must be able to show that there are important counterfactual supporting generalisations that cannot be expressed at a different level of analysis (e.g., functional or physical). Or conversely, 'if under a particular description of a system's behaviour, a physical, neural, or purely functional account captures all the relevant generalisations hence serves the explanatory function, then appealing to representations is not essential (ibid., p. 26).'[12]

Still, to claim that there is a special class of semantic-level counterfactual-supporting generalisations that characterise various cognitive

12 A similar formulation of the main motivation for postulating an independent semantic level is the following: 'The principle that leads us to postulate representational states (individuated by their content) that are distinct from functional states is exactly the same as the principle that leads us to postulate functional states that are distinct from physical states. In both cases we want to capture certain generalisations. We discover that in order to do this, we must adopt a new vocabulary and taxonomy; hence, we find ourselves positing a new, autonomous level of description' (Pylyshyn 1984, p. 32).

phenomena does not suffice to justify the claim that mental contents have an explanatory function in computational theories of cognition. Consider first some basic facts about the construction and evaluation of computational models in cognitive science. A preliminary condition for saying that a computational device/system models a particular cognitive capacity is that it satisfies an input-output or weak equivalence criterion. That is, the device must, under certain carefully specified conditions, yield the right type of output in response to an appropriate class of inputs. However, many authors have pointed out that this sort of behavioural evidence does not suffice to establish whether a certain computational model offers an adequate description or, more strongly, an explanation of a particular cognitive capacity. If it is to do the latter, the model must satisfy a stricter condition than mere behavioural mimicry, i.e., something along the lines of the *strong equivalence* criterion (cf. Pylyshyn 1984). The point of formulating such a stricter criterion is to guarantee that the computational modelling practice will avoid unprincipled and *ad hoc* moves:

> Psychologists pursue the goal of explanation, which means that, although we pursue a constructivist program, we must make sure our use of computer models is principled. This, in turn, means that stringent constraints must be applied to the theory construction task to ensure that the principles are separated from the ad hoc tailoring of the systems to make them fit the data. [...] we must make sure we specify the constants of the model in order to determine whether there are fewer degrees of freedom than are data points (ibid., p. 85).

The strong equivalence criterion is supposed to reflect the variety of strategies and techniques scientists use for improving both the empirical adequacy and explanatory power of current computational models of specific cognitive capacities.[13] The visual processing models discussed in section 2.2 illustrate several of these constraints which range from empirical or experimental-based constraints up to various theoretically-driven hypotheses. For instance, modelling the response of V1 cells in terms of Gabor patches or wavelets is driven both by results of simultaneous recordings from the LGN and V1 neurones and by higher-order principles like optimal visual redundancy reduction. Similarly, Jones et al.'s (1997) model of the top-down influences on early vision in terms of flexible models has been prompted by the idea of solving the tension between the poor performance of early visual systems and the high-quality of visual percepts.

13 As such, the satisfaction of this criterion is better seen as a matter of degree, which means that not even strong equivalence provides necessary and sufficient conditions for something counting as a definitive computational theory of cognition. This caution is particularly pertinent because, as it will be shown in what follows, the criterion remains incomplete in significant respects.

More generally, if a formally specified computational process is to be a serious candidate as an explanatory model of mental processing, then one should provide as explicit an account as possible of the way the model relates to the empirical phenomenon it is to explain. This in turn implies that one should seek to establish various constraints on the class of possible input-output equivalent computational mechanisms, e.g., by considering time and resource constraints as well as the capacity of such mechanisms to produce specific inferential and systematic patterns of behaviour. For instance, the 'qualitative' recognition model which postulates computations over ordinal relations between structural features of perceived images fares better with respect to such strong equivalence criteria than other models which construed the recognition problem in terms of comparisons of Gabor jet vectors or other related models (cf. Sinha and Balas 2008). More specifically, the computations postulated by the 'qualitative' recognition model do not raise special computational tractability issues that other previous models do, and the number of false positives generated by the model is close to insignificant. In addition, the model correctly predicts the patterns of successful recognitions and failures under a wide variety of illumination conditions. In light of these features, the model is a better candidate as an explanatory model of object recognition than other alternative models.

In summary, the practice of constructing adequate computational models of cognitive capacities seems to be constrained by a series of principles and norms, that include various higher-order or top-down considerations about general patterns such as the systematicity, compositionality, inferential coherence of certain cognitive phenomena such as thought and language. Whilst some authors have argued that the adoption of such constraints implies that the correct cognitive architecture of the brain must be a classical (symbols and rules) architecture (e.g., Fodor and Pylyshyn 1988), they have been criticised for failing to provide an unambiguous characterisation of the general (top-down) patterns that are supposed to entail such a strong conclusion (e.g., Smolensky 1988b; Matthews 1997; Frank, Haselager, and van Rooij 2009). One possible response strategy would be to insist that the difficulty of saying something more precise about these features stems precisely from their generality.

Another more compelling solution would be to take a closer look at the top-down principles that guide the actual scientific practices of developing better computational models/theories of cognition. Despite the fact that such a perspective is likely to yield a very fragmented picture of the aims, interests, and principles used by cognitive scientists in their modelling and explanatory practices, I claim that it nevertheless supports the general hypothesis that semantic interpretations (or the assignment of mental contents to the component parts of computational models) play an important normative role in

determining which of the proposed models provides an adequate explanation of the target cognitive capacity. This is essentially because the phenomena that are to be explained by a cognitive theory are typically characterised in broadly externalist (semantic terms). Whilst this observation does not imply an externalist (semantic) computational individuation strategy, it nevertheless vindicates the main insight of the semantic view of computationalism, which is expressed in the following quote:

> Although in the computational model, the symbolic codes themselves do not specify their intended interpretation (and the model's behaviour is not influenced by such an interpretation), the cognitive theory that makes claims about what it is a model of, which aspects of it are supposed to model something and which are not, *does* have to state what the states represent, for reasons already noted, having to do with explanation and with capturing generalisations. The cognitive theory would be gratuitous, or at best, weakly equivalent or 'mere mimicry,' if the ascription of some particular representational content were not *warranted*. The particular interpretation placed on the states, however, appears to be extrinsic to the model, inasmuch as the model would behave in exactly the same way if some other interpretation had been placed on them (Pylyshyn 1984, pp. 42-43).

I take this passage to support the separability of the individuation and explanation issues advocated in this chapter. Moreover, claiming that semantic interpretations are extrinsic to the models themselves does not commit one to any reckless form of subjectivism. For, as noted above, ascriptions of mental contents are constrained in a number of ways: they have to reflect the relevant top-down (higher-order) regularities and have to be assigned to appropriately typified structures. Thus, the defender of classical computationalism need not claim that semantic (externalist) contents play a role in the type-individuation of computational states in order to justify the important contributions that semantic interpretations make to the construction of adequate explanatory models of cognitive capacities.

A similar account of the role played by mental contents in computational theories of cognition can be found in Frances Egan (2010, 2013). The latter is taken to be an extension of her internalist account of computational individuation. By analysing her position, I seek both to clarify the main source of the internal tension threatening Egan's computationalist views and to provide a more adequate account of the role played by mathematical descriptions in computational cognitive explanation.

4.3.2 *Cognitive interpretations as gloss*

Egan (1999, 2010, 2013) acknowledges that the internalist view of computational individuation does not constitute a complete analysis of the practice of developing adequate computational models of cognitive phenomena. In addition, one must provide an account of how abstract computational models are connected with their target cognitive capacities. In order to do this, she maintains that one must say why and when computational theorists appeal to ascriptions of mental contents. Egan's proposal is that mental content ascriptions serve *several* important epistemic and pragmatic functions in the context of computational psychology. In particular, she claims that mental contents are essential in assessing the explanatory value of a computational theory of cognition.

In her 2010 article, Egan calls the ascription of mental contents to a formally characterised computational system: 'the cognitive interpretation of the model'. She nevertheless insists that this type of cognitive interpretation should be 'sharply distinguished from the mathematical interpretation specified by f_I. Only the latter plays an individuative role' (Egan 2010, p. 256). I have already shown that there are good reasons to resist the idea that mathematical functions play a proper individuative role with respect to computational systems/states. In what follows, I expand the previous critical analysis by discussing Egan's notion of *cognitive interpretation*. I then propose, in section 3.3, an alternative way of thinking about the roles played by the mathematical (canonical) description of a computational model/system.

Among the roles played by contents in computational theories of cognition, Egan (1999) identifies two that are common to different modelling techniques used in other physical sciences and one which seems to be unique/specific to the field of cognitive psychology. More specifically, she argues that a cognitive (semantic) characterisation of a computational model serves an expository or presentation function, 'explicating the formal account which might not itself be perspicuous'. A similar line is found in Chomsky (1995), who maintains that intentional characterisations of abstract computational models serve only as an informal presentation tool, contributing to the 'general motivation' of a computational theory. In addition, Egan (1999) points out that representational contents can also be instrumental in the elaboration of partial models or theories, since:

> a computational theorist may resort to characterising a computation partly by reference to features of some represented domain, hoping to supply the formal details (i.e., the theory) later. In the meantime, contents can serve a reference-fixing function allowing the theorist to refer to

> states yet to be given a precise formal characterisation'(Egan 1999, p. 182).

The characteristic function that mental contents are taken to play in computational psychology has to do with the fact that the questions that antecedently define the domain of a psychological theory are usually couched in semantic or intentional terms. A semantic or cognitive characterisation of the postulated computational processes is taken to enable the theory to adequately address these questions. In other words, the cognitive interpretation is said to play a *bridging* role between the pre-theoretical questions defining the psychological domain and the formal computational theory:

> The cognitive characterisation is essentially a *gloss* on the more precise account of the mechanism provided by the computational theory. It forms a bridge between the abstract, mathematical characterisation that constitutes the explanatory core of the theory and the intentionally characterised pre-theoretic explananda that define the theory's cognitive domain (Egan 2010, pp. 256-257).

On this view, in order to assess whether a particular computational model has an explanatory value with respect to a target cognitive capacity, the processes and structures postulated by the computational model need to be construed under a cognitive interpretation as representations of proximal or distal features of some external environment (for instance in the case of early visual processing, as edges, joint angles, etc. or in the case of visual recognition as invariant features of the object). Only when this is done, one can definitely say whether the model answers the questions that motivated the search for a computational theory in the first place. In consequence, the bridging role hypothesis claims that mental contents are fit to play only the epistemic role of connecting pre-theoretically or incompletely specified cognitive processes with formally specified models of these processes.

A *prima facie* difficulty with thinking of the roles played by mental contents in computational models along these lines is that it seems to make the interpretation of computational mechanisms look arbitrary and unprincipled. This would seem to follow because the idea that mental contents play a series of broadly epistemic roles is compatible with the fact that the semantic interpretation of any computational state/system is non-unique. However, as suggested above, this purported non-uniqueness of the cognitive interpretation does not necessarily amount to its being *ad-hoc* or non-objective:

> To call the cognitive characterisation a 'gloss' is not to suggest that the ascription of representational content is unprincipled. The posited states and structures are not interpretable as representations of distal visible properties

> (as, say, *object boundaries*, or *depth* or *surface orientation*) unless they co-vary with tokening of these properties in the subject's immediate environment (Egan 2010, p. 257).

This remark is in line with the observation made previously that the assignment of mental contents is actually guided by a series of norms and general principles which make semantic interpretations seem less arbitrary or indeterminate. Another way of thinking about this issue is to consider cognitive interpretations as being constrained by something like a directness requirement which reduces the class of possible content ascriptions. That is, given a certain cognitive modelling context, the ascription of certain distal features to the component structures of a specific computational model may well be the most straightforward (salient) and adequate interpretation. However, by itself, this requirement does not imply that the structures postulated in computational theories of cognition must necessarily represent their normal distal causes, for, in some cases, the computational structures postulated by a computational theory may equally well be interpreted as representations of proximal features (e.g., as discontinuities in the image, brightness ratio magnitude, etc.).

Thus, from a modelling perspective, both broad and narrow contents can play the role of creating a link between the semantically characterised explananda of computationalist cognitive theories and the abstract explanatory structures postulated by such theories. However, adopting a broadly externalist view with respect to the problem of cognitive computational explanation does not commit one also to an externalist view of computational individuation. Instead, I have argued that the correct view of computational individuation mandates that only formal properties/relations determine the type-identity of computational states/systems. Having established that externalist (semantic) contents do not count as criteria for computational individuation, the only remaining point that requires clarification concerns the relationship between the mathematical canonical description and the so-called cognitive interpretation.

On Egan's own account this relationship turns out to be quite problematic for two reasons. The first is that Egan (2010) claims both that the mathematical description constitutes 'the explanatory core of the [computational] theory' and also that mental contents play the essential role in the evaluation of the explanatory value of candidate computational models of cognition. One possible way to read the two claims so that they are not in tension with one another is to say that the mathematical description constitutes the *explanans* of the computational model/theory, which then is cast in an appropriate cognitive interpretation. But even under this reading it is unclear whether it is the mathematical description itself that does the explanatory job or rather the cognitive interpretation. Therefore, Egan seems to face the challenge of giving a more detailed account of how is it that mathe-

matical structures can play a proper explanatory role in an empirical theory of cognition.

The second reason why this relationship is problematic has to do with the status of mathematical functions (descriptions) themselves. On Egan's own account, mathematical structures are sometimes said to play a genuine explanatory role in computational cognitive theories (as per the previous quote), and, at other times, they are taken to play an essential role in individuating computational states/systems. Claiming that mathematical structures (functions) serve both to individuate computational states/systems and to explain specific cognitive capacities or patterns risks conflating the individuation and explanation issues all over again. In line with an internalist (formal) view of computational individuation, I maintain that mathematical descriptions are better viewed as playing an *explanatory* role in computationalist theories of cognition.

4.3.3 *The structure of classical computationalist explanations*

In light of the previous arguments, I claim that classical computational explanations of cognitive capacities proceed by decomposing a complex cognitive task (e.g., object recognition) into a set of more basic tasks (e.g., edge extractions, feature construction, ordinal matching, etc.) which in turn are characterised in terms of a series of computing operations defined over appropriately typified symbols. Thus, computational explanations connect, via a number of identifiable steps, the target cognitive phenomenon/pattern (which is often characterised in semantic or intentional terms) to a computational structure which reveals certain fundamental features of the cognitive capacity in question. These features contribute to a better understanding of the cognitive phenomena under investigation by allowing cognitive scientists to test a wider range of counterfactual generalisations pertaining to the target cognitive phenomena, and to draw connections between the computational descriptions of what sometimes seem to be different types of cognitive structures.

I have argued that the assignment of semantic (broadly externalist) contents plays a crucial role in the development of these computational models because it serves to justify why a particular computational structure/system can be taken to capture something relevant about the structure of the target cognitive phenomenon. In addition, I have shown that in order to play this sort of normative role in the construction and refinement of good explanatory models of cognitive phenomena, the interpretation function need not be taken to be a one-to-one mapping between computational states and semantic contents. In fact, even a cursory glance at the modelling practices of cognitive scientists shows that semantic interpretations of computational mod-

els are most of the time partial and mixed (i.e., comprising both what philosophers have identified as broad and narrow mental contents).

Besides semantic interpretations, the explanatory adequacy of particular computational models is usually evaluated by taking into account additional constraints, developed along the lines of the strong equivalence criterion (Pylyshyn 1984). According to the latter, the input-output (or weak) equivalence of the cognitive system and the computational system/model may not suffice to guarantee that the proposed model has a genuine explanatory value. In addition, modellers take into consideration quantitative measures such as response times, complexity profiles of the modelled and modelling systems, and so on, which further increase the adequacy of the computational structures postulated by classical computational models/theories of cognition.

Furthermore, I claim that taking semantic interpretations to function as norms which guide the construction of good computational theories of cognition is compatible with the idea that the mathematical (canonical) descriptions of computational theories/models of cognitive capacities play an explanatory rather than individuative role. Mathematical descriptions can be taken to *identify* or index the computational structures postulated by particular models/theories of cognitive capacities. Identifying computational systems in terms of the mathematical functions they compute is theoretically useful since it makes these (often highly complex) systems theoretically tractable by identifying the variables of a system that are relevant from a modelling point of view.

Since these mathematical descriptions identify certain stable (more fundamental) features in the cognitive phenomena being investigated, which in turn support relevant counterfactual generalisations and provide further insight about the structure of the target cognitive phenomena, they can be said to play an explanatory role in these scientific practices. This way of thinking about the role of mathematical descriptions in computationalist modelling is also consistent with the discussion of the contributions of mathematisation to the dynamic systems modelling practices analysed in chapter 3. In concluding this investigation, I would like to emphasise a number of advantages of adopting this multilayered view of computational explanations of cognition.

4.4 CONCLUDING REMARKS

The interpretation of classical computationalism proposed in this chapter comprises two main hypotheses. First, the individuation hypothesis claims that the internal states and structures of a computational model are type-identified on purely formal grounds, i.e., in virtue of their structural (syntactic) properties. This hypothesis is in line with

the formality constraint endorsed by virtually all classical accounts of computation; but it contravenes the orthodox reading of the semantic view of computation, according to which semantic (externalist) principles and/or contents play an essential role in computational individuation. In other words, in contrast to the latter claim, the formal individuation thesis denies that mental contents (both broad and narrow) impact the individuation of the internal states of computational systems. Second, the broadly externalist explanation hypothesis that constitutes the other pole of classical computationalism rescues mental contents from a strictly eliminativist picture of computational explanation. I have argued that the semantic (cognitive) interpretation of the formally specified structures postulated by computational models of cognitive phenomena plays a normative role in establishing the adequacy and explanatory value of such a computational model.

Therefore, the account being put forward has an internalist component because it claims that the *individuation* of computational systems/states is affected only by structural features (properties, relations) of their component parts. However, as a general account of computationalist *explanation*, the proposed view is compatible with a weak form of semantic externalism which acknowledges that the practice of constructing adequate computational models of cognitive phenomena is constrained by the semantic interpretations that can be assigned to these models. This further implies that assignments of mental contents (broad and narrow) play a substantive role in the construction and evaluation of the explanatory value of particular computational models of cognitive phenomena. That is, they constitute a bridge between the explananda of cognitive theories/models and the abstract (mathematical) explanatory structures postulated by classical computational models/theories.

The analysis carried out in this chapter was intended to identify the main theoretical principles which underlie classical computationalist approaches to cognition. I have characterised the explanatory scheme associated with classical computationalism as a decompositional strategy which elucidates particular cognitive phenomena by revealing in a stepwise fashion certain more fundamental abstract features of the phenomena being investigated. For the purposes of this analysis I have relied broadly on the classical Marrian picture of computational explanations (Marr 1982). From a theoretical point of view, Marr's tripartite conception proves to be very helpful in allowing a clear articulation of the internalist (formal) view of computational individuation. Moreover, I think that this picture provides a convenient way of characterising the different levels of abstraction at which cognitive phenomena can be analysed.

However, from a practice-based perspective, the tripartite picture is better viewed as a simplification of a continuum of abstract models. As was shown in section 2.3, computationalist models of cognitive ca-

pacities are specified at multiple levels of abstraction and/or analysis. This observation further reinforces the idea that classical computationalist models are not developed in isolation from considerations pertaining to lower-levels of analysis. In fact, Marr (1982) himself insisted that no single level of description/explanation can be well understood without reference to the other levels. The present account recognises the importance of these factors which guide the construction of adequate explanatory models of cognitive phenomena.

Finally, I would like to comment briefly on a thorny issue often raised in debates concerning classical computationalism, namely the realist commitments entailed by this theoretical position. Throughout the investigations carried out in this chapter, I have explicitly ignored the fact that most supporters of classical computationalism are dyed-in-the-wool realists about mental computations. The main reason for doing this is that the almost exclusive focus on these realist concerns has driven too many philosophers to conflate the individuation and explanation issues which in turn generated a series of difficulties for the formulation of an adequate account of classical computationalist explanations of cognitive phenomena.

Against this trend, I have focused on the explanatory framework associated with classical computationalism and articulated an account which clarifies the main factors that contribute to the construction of adequate computationalist models/theories of cognitive phenomena. Within this setup, I have argued that the appeal to theoretical posits such as internal representations and rules is a preliminary condition for the scientific practice of developing computational models/theories of cognitive phenomena. There is however a mild realist concession that can be derived from the analysis of the problem of cognitive explanation. For the success of some of the proposed computational models/theories of cognition may be taken to provide some sort of vindication of the ontological principles presumed in these activities. However, this concession should not be taken to imply that the explanatory value of a scientific model/theory reduces to its existential commitments or that the latter by themselves guarantee the explanatory import of particular cognitive theories. I will pursue these themes further in the next chapter where I will analyse another model of cognitive explanation which combines the insights afforded by the classical computationalist and the mechanistic frameworks.

5

THE MECHANISTIC VIEW OF COMPUTATIONAL EXPLANATION

5.1 INTRODUCTION

The issues facing classical computationalism have encouraged the development of a series of alternative models of computational explanation. This chapter explores the attempt to extend the notion of computational explanation by incorporating it into the mechanistic account of computation defended by Piccinini (2007b, 2008a, 2010), Craver and Piccinini (2011), Milkowski (2010, 2013), and Kaplan (2011), among others. The *mechanistic conception of computational explanation* comprises two major hypotheses: (a) the functional view of computational individuation, and (b) the mechanistic view of computational explanation. According to the functional individuation hypothesis, computational states and processes are individuated in terms of their functional properties, i.e., non-semantically. More specifically, the functional properties that are taken to type-individuate computations are specified via mechanistic decompositions of complex (physical) systems. The individuation hypothesis is complemented by the mechanistic explanation hypothesis, according to which computational explanations are a special sub-class of mechanistic explanations.

5.1.1 *A motivational strategy for the mechanistic view*

There are three principal motivations that drive the mechanistic view of computation. Firstly, computational mechanists argue that the account constitutes a better alternative to classical computationalism, avoiding the pitfalls of the orthodox semantic view of computational individuation (cf. Piccinini 2008a). Defenders of mechanism claim that their proposed computational individuation scheme is neither context- or observer-dependent, nor does it rely on purported referential or ideological resemblances between the internal states and structures of a computational system. In other words, the functional view of computational individuation is supposed to deliver determinate and objective computational taxonomies which are used to type-identify specific computational devices.

Secondly, proponents of the mechanistic view claim that the account provides a more robust picture of both computational individuation and computational explanation that covers a wide range of experimental and theoretical practices, from computer science and engineering to computational psychology and neuroscience. And, thirdly, these theorists claim that their approach yields a sharp separation between the questions raised by computationalist approaches to cognition and those specific to theories of mental content. This in turn is supposed to help disentangle a number of philosophical debates that have systematically conflated the two issues.

In addition, they argue that the mechanistic view exhibits a number of desirable meta-properties which recommend it as a general encompassing picture of computation. Proponents of the account maintain that mechanism provides an objective picture of what it is for a physical system to be a computing device in the first place. Furthermore, the mechanistic view is said to support a robust distinction between computing and non-computing mechanisms, viz. by showing that the 'right' things compute (e.g., digital computers, calculators, etc.), whereas the 'wrong' things do not (e.g., the weather, planetary systems, etc.).

Lastly, it has been argued that the mechanistic view entails an appropriate account of the applicability of computability notions and principles to the study of cognition. More specifically, as an account of the applicability of the theory of computation to cognitive phenomena, the mechanistic view has two potentially interesting entailments. First, the account reinforces the useful distinction between computationalist models and computationalist explanations of cognitive capacities. Second, the functional individuation scheme promoted by the mechanistic account is arguably compatible with the actual taxonomies used by practicing cognitive scientists.

5.1.2 *Aims and outline of the argument*

Despite its promising features, the mechanistic account faces a number of challenges that undermine its claim to ascendancy over other versions of computationalism. In this chapter, I show that the mechanistic account actually conflates the individuation and explanation issues by equating computational individuation criteria with pragmatic explanatory principles. The mechanistic account also seems to distort some of the specific aims and purposes of computationalist approaches to cognition. This chapter therefore challenges the idea that the mechanistic view provides the most appropriate account of the applicability of computational models to the study of cognition. More importantly, I maintain that the mechanistic component of this view of computationalist explanation tends to undermine the very

idea of there being a distinct class of computational explanations of cognitive phenomena.

This chapter comprises four parts. I begin by analysing the conception of computational mechanism articulated in a series of papers by Gualtiero Piccinini (2007a, 2008a, 2010). Then, in section 3, I show the consequences of adopting his notion of generic computation for the issue of computational individuation. Section 4 explores the details of the mechanistic picture of computational explanation and the idea that the mechanistic framework facilitates the unification of various explanations of cognitive phenomena developed at different levels of analysis or abstraction. In the last section, I discuss some of the main limitations of the mechanistic account of computation, and conclude with a comparison between mechanistic and classical computationalism.

5.2 COMPUTING MECHANISMS

As stated in the introduction, the mechanistic view comprises two allegedly separable hypotheses pertinent in this context: (i) a functional hypothesis of computational individuation and (ii) a mechanistic hypothesis of computational explanation. I claim that both hypotheses can be viewed as consequences of the notion of generic computation developed on the mechanistic approach. By analysing this notion, I seek to identify the similarities between classical computationalism and mechanistic computationalism. I maintain that acknowledging the features that the two views have in common allows a better assessment of the distinctive contributions of each framework to the study of cognitive phenomena.

There are several distinct threads which contribute to the mechanistic conception of computation (cf. Piccinini 2007a, 2008a, 2010; Craver and Piccinini 2011; Piccinini and Bahar 2013). In particular, mechanists insist on the centrality of the distinction between abstract and concrete computation. Within the latter category they further distinguish between analog, digital and neural computation, and argue that only the notion of neural computation is an appropriate tool for elucidating the nature and structure of cognition (cf. Piccinini 2009; Piccinini and Bahar 2013). In what follows, I analyse the various distinctions introduced by the mechanistic account of computation in order to clarify how the notion of generic computation is related to that of neural computation, which in turn is taken to play a crucial role in computationalist approaches to cognition. The following discussion is also intended to sharpen certain ideas presented in chapter 4 concerning the relation between abstract and concrete computation.

5.2.1 *Abstract computation*

The abstract notion of computation has its origins in the pioneering work of Alan Turing (1937), Alonso Church (1936) and other mathematicians who sought to define in rigorous technical terms the intuitive notion of computable function. Their work in mathematical logic yielded a series of formally equivalent characterisations of the class of effectively computable functions. The results that came to be known as the Church-Turing thesis synthesise these efforts in the claim that all effectively computable functions are Turing computable (or, equivalently, recursive functions or abacus computable functions). Although the thesis is not obvious nor can it be rigorously proved (since the notion of effective computability is itself an intuitive and not a rigorously defined one), an enormous amount of evidence has been accumulated in support of it. What follows is a rough picture of the key ideas that underlie the mathematical work associated with the development of the technical notion of computable function.

From a formal point of view, to define a computation amounts to specifying a string of letters from a finite vocabulary and a list of instructions for generating new strings from old strings. An ordered list of instructions constitutes an algorithm, i.e., the abstract characterisation of a program. Thus, given a vocabulary (V) and a list of instructions appropriately defined over strings of V-letters, a computation is a sequence of strings such that each member of the sequence is derived from another member via some instruction in the list. Letters and strings are usually called symbols or symbolic structures in virtue of the fact that they are typically assigned semantic interpretations. However, as noted in the previous chapter, this does not imply that formally typified symbols possess any of their potential contents essentially. Thus, symbols are formally identified entities that may be concatenated to other symbols to form lists called strings. This in turn implies that strings which are complex symbolic structures are formally individuated solely by the types of symbols that compose them and their order within the strings.

This approximate characterisation of what counts as a computation is consistent with the idea that most interesting computations depend not only on the input strings of data but also on the internal state of the system that is said to perform the computation. Since internal states may also be defined as strings of symbols, the individuation criteria for input strings will apply to internal states as well. According to this picture, a computation comprises an initial internal state of the system together with an input string, a series of intermediate strings, and a final string consisting of the output plus the final internal state of the system. In addition, for all V-strings and any appropriately defined algorithm, it is possible to specify a general rule which characterises the function computed by the system that acts

in accordance with the algorithm (e.g., the addition rule). This rule offers a general characterisation of the relation which holds between inputs, internal states, and outputs produced by the system following the steps (instructions) of the algorithm. There are two important features which characterise this sort of rule, viz: (i) it is highly general, i.e., it connects all inputs and outputs from a relevant class; and, (ii) it is input-specific, in that it depends on the composition of the input for its application. Note that this type of rule is typically more abstract than any specific algorithm, in the sense that the same rule can characterise different algorithms, but it may also be equivalent to the algorithm itself.

The abstract conception of computation makes it clear that computations are type-identified only by the type of entity over which they are defined together with the algorithms (list of instructions) that specify how these entities are manipulated and transformed to yield other types of entities. However, proponents of the mechanistic view insist that the notion of abstract computation does not suffice to elucidate either the conceptual or the empirical implications of computationalist approaches to cognition. They claim that the mathematical notion of computation applies directly only to abstract systems/mechanisms. Mechanists therefore require, in addition, a clarification of how this abstract notion applies to concrete (physical) mechanisms.

5.2.2 *The varieties of concrete computation*

In order to grasp the connection between the mathematical notion of computation and the varieties of concrete or physical computation discussed in the scientific and philosophical literature, it is helpful to note that the abstract notion provides a sketch for a broad or generic conception of computation. Digital, analog, and neural computation will in turn be shown to be species (among others) of this generic view of computation. One of the clearest mechanistic formulations of the generic notion of computation says that:

> Computation in the generic sense is the processing of vehicles (defined as entities or variables that can change state) in accordance with rules that are sensitive to certain vehicle properties and, specifically, to differences between different portions (i.e., spatiotemporal parts) of the vehicles. A rule in the present sense is just a map from inputs to outputs; it need not be represented within the computing system. Processing is performed by a functionally organised mechanism, that is, a mechanism whose components are functionally organised to process their vehicles in accordance with the relevant rules (Piccinini and Bahar 2013, p. 458).

On this definition of generic concrete computation, the properties (of the vehicles and instructions) that are counted as relevant for the purposes of computing are independent of the physical media that implement them. More specifically, mechanists claim that computational structures (i.e., the vehicles) are medium independent in the sense that the input-output map (the general rule) that defines a computation 'is sensitive only to differences between portions of the vehicles along specific dimensions of variation - it is insensitive to any more concrete physical properties of the vehicles' (Piccinini and Bahar 2013). Otherwise put, the general rules that define specific computational processes are 'functions of state variables associated with a set of functionally relevant degrees of freedom, which can be implemented differently in different physical media' (ibid.). This implies that a particular computation can be implemented in multiple physical media (e.g., mechanical, electro-mechanical, electronic, etc.) provided that the candidate implementational bases have enough degrees of freedom (i.e., distinguishable states) that can be appropriately accessed and transformed. Defenders of the mechanistic position claim that, despite its permissiveness, this notion of generic computation rules out a host of physical mechanisms as being non-computational (e.g., stomachs, the weather, planetary systems, the Watt governor, etc.).

It should be noted that nothing that has been said so far about the notion of generic computation contravenes to the internalist (formal) view of computational individuation defended in chapter 4. However, mechanists seem to insist that one major advantage of their way of thinking about computation is that it rules out upfront the semantic view of computational individuation by showing that computations can be defined in a purely formal way, independently of any semantic characterisation/interpretation. In addition, the generic notion of computation is said to allow the clear differentiation of three more specific notions of computation: digital, analog, and neural computation. Among these, mechanists maintain that only neural computations are appropriate devices for investigating the structure of cognitive phenomena. I begin by analysing the mechanist formulations of the notions of digital computation and analog computation, before considering the notion of neural computation itself.

5.2.2.1 *Digital computations*

Perhaps the most popular notion of computation discussed both in the computer scientific and philosophical literature is that of digital computation. As with all the other notions of computation analysed in this section, that of digital computation admits of both an abstract and a concrete or physical characterisation. At the abstract level, digital computation can be characterised along the lines of Turing's original proposal sketched above, viz. as the manipulation of strings of

discrete elements (symbols) (cf. Turing 1948/2004). In order to obtain a concrete characterisation of the notion of digital computation, the first step is to establish the concrete counterpart of the formal notion of symbol. Mechanists have suggested that symbolic structures postulated at the abstract level of defining a computing system may be physically implemented by 'digits'. A digit is a macroscopic state (of a component of a physical system) whose type can be reliably and unambiguously distinguished by the system from other macroscopic state types. Just as their abstract counterparts (viz. symbols), digits can be ordered to form sequences of strings which in turn serve as the vehicles of computation. Thus, a concrete digital computation is defined as the processing of strings of digits in accordance with a rule, which is simply a map from input strings and internal states, to output strings. Although they qualify as concrete computations, operations over strings of digits are not responsive to any specific kind of physical property. This seems to imply that digital computations can be implemented by any physical medium with the right internal composition and organisation. In summary, the concrete conception of digital computation entails that:

> Digits are unambiguously distinguishable by the processing mechanism under normal operating conditions. Strings of digits are sequences of digits, that is, digits such that the system can distinguish different members of the set depending on where they lie along the string. The rules defining digital computations are, in turn, defined in terms of strings of digits and internal states of the system, which are simply states that the system can distinguish from one another. No further physical properties of a physical medium are relevant to whether they implement digital computations (Piccinini and Bahar 2013, p. 459).

Mechanists claim that defining digital computation in terms of operations over digits affords a more general conception than three other related and more commonly invoked notions: classical computation (cf. Fodor and Pylyshyn 1988), algorithmic computation, and Turing-computable functions. However, this claim seems to be overstated. Firstly, as I have shown in the previous chapter, classical computationalism need not be committed to the hypothesis that all digital computations are defined exclusively over language-like vehicles and thus need not be seen as more restrictive than the present conception of computation. Secondly, the motivations for saying that concrete digital computations go beyond Turing-computability are not entirely clear, especially since mechanists tend to be quite sceptical about the notion of hypercomputation (e.g., Copeland 2002; Copeland and Shagrir 2011; Piccinini 2007b). And, lastly, as suggested above, the notion of algorithm is just another variant of describing the input-output mapping that connects the input strings of digits with the output

strings of digits, and thus again cannot be said to yield a more restrictive notion of computation.

Although I do not think that mechanists can make a strong case for any of these differences, I concede that the mechanistic formulation of the notion of digital computation has the advantage of avoiding certain misleading implications that are typically associated with the other proposed accounts. In particular, since on the mechanist framework the notion of digital computation is presented explicitly as an extension of the abstract notion of Turing-computation, this formulation avoids the temptation of claiming that strings of digits have essential semantic interpretations (contents). That is, when considered outside any specific modelling context, the formal characterisation of digital computation becomes the most salient.

Thus, to recapitulate, according to mechanists, a physical system counts as a digital computing system to the extent that it is 'functionally organised to manipulate input strings of digits, depending on the digits' type and their location on the string, in accordance with a rule defined over the strings' (cf. Piccinini and Bahar 2013, p. 460). This characterisation implies that there are two distinctive features of the notion of digital computation:

> (a) whether a particular microscopic state belongs to a digit type is unambiguous relative to the behaviour of the system; and (b) the output of the computation depends (either deterministically or probabilistically) only on the internal state of the system and on the number of input digits of their types, and the way they are concatenated within the string during a given time interval (ibid., p. 460).

From this definition it follows that a wide range of physical systems qualify as performing digital computations: Turing machines, finite state automata, ordinary computers and calculators, and perhaps more surprisingly certain physical implementations of connectionist networks such as perceptrons (e.g., Minsky and Papert 1972) and McCulloch-Pitts nets (McCulloch and Pitts 1943). Whilst this classification seems to support the claim that the mechanist definition of digital computation is more general than the one typically invoked in discussions of classical computationalism, I submit that this generality is primarily a consequence of the fact that mechanists characterise the notion of digital computation in a way that is independent of the particular tasks or problems that such computational systems/devices are taken to solve or model.

Whilst the notion of digital computation (both abstract and concrete) is the notion that can be credited with inspiring the computational theory of cognition, various authors have claimed that the internal dynamics of cognitive processes points towards a different notion of computation. One such candidate, analog computation, is

analysed in some detail by proponents of the mechanistic view, but for reasons that I will spell out below, it is found to be inappropriate for the purposes of modelling and explaining cognitive phenomena.

5.2.2.2 *Analog computations*

The notion of analog computation is more difficult to pin down than that of digital computation. For instance, according to one very broad sense of the term 'analog', a process qualifies as analog if it can be characterised as the dynamical evolution of real variables in real time. At least some of the authors who have proposed that the notion of analog computation is adequate for the study of cognitive processes employ precisely this conception (cf. Churchland and Sejnowski 1992; van Gelder 1998). The problem with the proposal is that most systems count as analog in this sense, even those that intuitively are not proper computing systems, such as the planetary motions, digestion, and so on.

Another sense of the term 'analog' encountered in the cognitive literature refers to representations that are *analogous* with what is being represented. For instance, certain models of visual (Hubel and Wiesel 1962) and auditory (Schreiner and Winer 2007) receptive fields make reference to such analog representations constructed by the relevant parts of the nervous system. However, as some mechanists have pointed out, this loose talk of analog models does not suffice to establish whether there are certain brain functions adequately described in terms of analog computations (e.g., Piccinini 2008b). Despite its limitations, a number of authors (cf. Pour-El 1974; Rubel 1985, 1993; Mills 2008) have argued that analog computation provides an alternative, better, set of principles than digital computation does for the study of cognitive phenomena. For this reason, it is worth clarifying further the main differences that separate the two conceptions.

Mechanists claim that the distinction between analog and digital computation consists in the fact that physically implemented continuous variables are quite different entities from the strings of digits over which digital computations are defined. For one thing, a digital computer can distinguish between different types of digits in an unambiguous manner, whereas a physically implemented analog computer cannot do the same with the exact values of continues variables, simply because the latter can be measured only within a certain margin of error. It is primarily in virtue of this feature that mechanists (e.g., Piccinini 2007b, 2008a,b; Piccinini and Bahar 2013) maintain that analog computing systems constitute a different class from digital computers. Nevertheless, analog computation has an important feature emphasised in the case of generic computation, namely that of being defined over environment independent entities (vehicles). This is important because it implies that, just like other types of computa-

tions, analog computations are type-individuated in principle solely by their formal or structural properties.

Turning to analog computationalism as a hypothesis about biological cognitive systems, it has been claimed that functionally significant signals manipulated by the nervous system are irreducibly continuous variables. Thus, there seems to be a *prima facie* case for the analog computationalist hypothesis in the fact that both types of systems (i.e., analog computers and brains) have a continuous dynamic. Neural inputs - viz. neurotransmitters and hormones - are most usefully modelled as continuous variables and their release and uptake is modulated by chemical receptors that operate continuously in real time. Similarly, dendrites and at least some axons transmit graded potentials, i.e., continuously varying voltage changes. These and other features seem to support the comparison between brains and analog computers. However, in addition to these preliminary similarities, there are also several major differences between nervous systems and analog computing systems.

For instance, while it is true that the firing threshold of individual neurones is subject to modulation by continuous variables (such as ion concentrations), graded potentials vary continuously, and spike timing may be functionally relevant, none of these aspects of neural signals are similar enough to the components of analog computing systems to allow the application of the mathematical theory of analog computation to the understanding of brain functions. More importantly, in the case of spikes which are presently thought to be functionally the most significant signals transmitted by neurones, it is not the absolute value of the voltage which is treated as functionally relevant, but rather the fact of whether a spike is present or not. As a consequence, spikes are said to have an all-or-none character and neuroscience presently focuses on firing rates and spike timing as the principal functional variables at the level of neural networks. This generates a disanalogy between nervous systems and analog computing systems because there does not seem to be anything resembling firing rates or spike timing in the architecture of analog computers. All these remarks seem to weaken the main claim of analog computationalism, namely that brains are analog computers.

Although there are other notions of computation that can be defined as subspecies of the generic notion of computation (e.g., quantum computation), mechanists insist that the notion which is most pertinent in the context of cognitive psychology and neuroscience is that of *neural computation*. They claim that neural computation constitutes a *sui generis* type of computation which is indispensable for properly assessing the explanatory value of proper computational models/theories of cognition.

5.2.2.3 *Neural computations*

The mechanistic argument for postulating neural computations as a distinct species of generic computation is grounded in two claims: (i) the neurally relevant units of computation cannot be either digits or strings of digits; and (ii) the neuroscientific community makes use of a notion of neural computation which satisfies the principal mechanistic criteria for generic computation. Recall that according to the mechanistic picture, something counts as a generic computational process if and only if it can be defined over environment-independent vehicles which are manipulated and transformed in accordance with a general, input-sensitive rule. Mechanists hold that neural computations can be characterised in terms of these two parameters without assuming that the variables processed by neural computations are anything like digits or strings of digits (i.e., the vehicles of digital computation).

The mechanistic notion of neural computation is best seen as a direct response to the computationalist hypothesis defended by McCulloch and Pitts (1943), according to which neural processes can be appropriately conceived as digital computations performed over spikes. The principal empirical justification proposed for this hypothesis is the alleged similarity between digits and spikes, viz. spikes appear to be discrete or digital, that is, they are unambiguously typified functional units of a cognitive neural system. To this, McCulloch and Pitts (ibid.) added the assumption that sets of spikes are strings of digits, which in turn requires that spikes be concatenated. There are at least two ways in which to conceive of this concatenation relation. In the case of spike trains from a single neurone, the obvious candidate for the concatenation relation consists in the temporal ordering of the spikes. In the case of classes of spikes from different neurones occurring within well-defined time intervals, a concatenation relation might be defined by first identifying a relevant set of neurones and then taking all spikes occurring within the same time interval as belonging to the same string.

In response to this traditional hypothesis, mechanists argue that 'the suggestive analogy between spikes and digits - based on the all-or-none character of spikes - is far from sufficient to treat a spike as a digit, and even less sufficient to treat a set of spikes as a string of digits' (cf. Piccinini and Bahar 2013, p. 468). The argument against McCulloch and Pitts (1943) digital conception of neural computation divides into two steps which basically deny that: (i) spikes can be appropriately typed as digits and (ii) spikes can be fitted into strings of digits. Mechanists begin by denying the founding assumption of the traditional conception of neural computation according to which spikes are digits. They point out that since digits belong to finitely many types which are 'unambiguously distinguishable by the system that manipulates them', for spikes (and their absence) to be digits, it

must be possible to individuate them into finitely many types that are equally unambiguously distinguishable by neural mechanisms. However, this proposal is highly problematic for a number of reasons.[1]

To get their argument off the ground, mechanists endorse the assumption that the most functionally significant variables for the purposes of understanding neural computational processes are the properties of neural spike trains such as firing rate and spike timing. With this hypothesis in hand, they argue that the neuroscientific evidence concerning these sorts of properties supports the conclusion that neither spike trains from single neurones nor sets of synchronous spikes from multiple neurones are viable candidates for strings of digits. Since without strings it is not possible to define appropriate operations over strings, mechanists claim that the digital conception of neural computation is bankrupt. Instead, they propose that neural computation is *sui generis*:

> In a nutshell, current evidence indicates that typical neural signals, such as spike trains are graded like continuous signals but are constituted by discrete functional elements (spikes). Therefore, typical neural signals are neither continuous signals nor strings of digits; neural computation is *sui generis* (cf. Piccinini and Bahar 2013, p. 477).

In further support of their contention, mechanists also point out that theoretical neuroscientists build mathematical models of neural mechanisms and processes in which 'the explanatory role of (digital) computability theory and (digital) computer design is nil' (ibid.). They insist that understanding neural computation requires specially designed mathematical tools rather than the mathematics of digital and analog computation. More generally, the mechanistic argument is taken to challenge 'not only classical computationalism, which is explicitly committed to digital computation, but also any form of connectionist or neurocomputational theory that is either explicitly or implicitly committed to the thesis that neural activity is digital computation' (cf. ibid., p. 469). So, whilst the notion of digital computation sketched above is taken to constitute a distinctive sub-species of generic computation, mechanists challenge the idea that this notion is appropriate in the context of modelling and explaining cognitive phenomena.

Their criticism, which is primarily targeted at the traditional picture of neural computation (e.g., McCulloch and Pitts 1943), has been taken to have three important consequences for the debates concerning computational approaches to cognition. Firstly, the notion of generic computation, of which neural computation is only a subspecies, is taken to yield a functional computational individuation strategy that goes hand in hand with a mechanistic conception of

1 For a detailed discussion, see Piccinini and Bahar (2013, pp. 469-474).

computational explanation. Secondly, the focus on the notion of neural computation is meant to bring out how computational modelling of cognitive capacities differs as an activity from other applications of computability theory to the empirical domain. And, thirdly, the mechanistic arguments are supposed to bridge the gap between cognitive psychology and neurobiology, by providing a framework in which the different taxonomies and explanatory strategies used in these fields can be appropriately integrated. In what follows, I analyse these three key consequences of the mechanistic view of computational explanation and their connections with the account of classical computationalist explanations developed in chapter 4.

5.3 THE FUNCTIONAL VIEW OF COMPUTATIONAL INDIVIDUATION

As stated above, both the computational individuation and explanation hypotheses can be seen as consequences of the mechanistic notion of generic computation. I start by analysing the specific version of the individuation hypothesis proposed by supporters of mechanism, viz. *the functional view of computational individuation*. For mechanists, this picture of computational individuation is supposed to hold for all sorts of computational models used in different branches of science, including among others, cognitive psychology and neuroscience (cf. Piccinini 2007a, 2008a; Craver and Piccinini 2011). More specifically, mechanists maintain that scientific practice across a large number of domains supports a *wide functional individuation* strategy. I analyse the arguments supporting this individuation hypothesis as well as the claim that wide functional individuation is distinct from computational individuation by wide functional contents. The latter point is important for establishing whether or not the mechanist hypothesis of computational individuation counts as a semantic individuation strategy.

There are two main arguments for the functional view of computational individuation. One basically consists in the critique of the semantic view of computational individuation. Given that I have argued at length in the previous chapter against the adoption of a semantic view of computational individuation, I will not elaborate on these objections further. Whilst I agree with most of what the mechanists have to say against the semantic individuation hypothesis, I have also shown that it is possible to reconstruct the classical computationalist position so that it no longer entails a semantic individuation scheme. This in turn implies that the outcome of criticising the semantic computational individuation strategy does not necessarily amount to a wholesale rejection of classical computationalism.

The other major argument for the functional view of computational individuation rests on the definition of generic computation introduced in the previous section. Computation in the generic sense is

defined as the processing of specific types of medium-independent entities (i.e., vehicles) in accordance with general rules that are sensitive only to some of the properties of these entities.[2] Since the vehicles of generic computations are said to be medium-independent, it follows that the type-individuation of particular computing systems/structures is determined solely by certain structural (formal) features of their constituents. Because these individuative features are distinct from the detailed physical descriptions of the computational systems of interest (e.g., program-controlled computers, analog computers, brains, etc.), mechanists propose to call them *functional* properties (cf. Piccinini 2004, 2007a, 2008a,b). This label also serves to convey the idea that the mechanistic computational individuation hypothesis should yield a genuinely *non-semantic* taxonomy. The key idea of the functional view of computational individuation is that the internal states and structures of a computational model (mechanism) are individuated solely in terms of (a subclass of) their functional properties.

As suggested above, the functional view of computational individuation is meant to cover all types of computational systems, both abstract and concrete. For instance, when applied to the case of standard Turing machines (TM), this individuation strategy characterises TMs in terms of two principal components: (1) a (potentially infinite) tape, whose function is to hold symbols (letters), and (2) a scanner whose function is to move along the tape, writing or erasing symbols on it. Particular TMs are individuated by a finite list of conditional instructions (IFs) such as: if the current scanned cell of the tape contains a particular type of symbol and the scanner is in a certain state, then the scanner prints a particular letter, moves one step to the left (or right) and goes into a new state. Therefore, each particular TM is uniquely individuated by its characteristic list of instructions, which also comes with an implicit appropriate vocabulary. The crucial point is that although these individuating descriptions of particular TMs might be semantically interpreted, they need not be in order to do their individuative job. Otherwise put, they do not essentially involve any semantic properties of the internal states of the device; rather, they characterise the functioning of the computing device, i.e., how the scanner, in virtue of its current inner state, sequentially changes or otherwise manipulates the symbol on the scanned part of the tape. The computational identity of specific TMs is fixed by their instructions, and not by the interpretations that may or may not be assigned to their inputs, outputs, and internal operations.

According to the mechanistic account, the same sort of individuation strategy applies to the whole range of computational systems, in-

2 As I have previously argued, this general rule can be an abstract characterisation of the algorithm or ordered set of instructions followed by the computing system or the algorithm itself. If the former, then the rule need to be explicitly represented in the computational architecture of the system.

cluding universal TMs, program-controlled computers, artificial neural networks, and neural computation (cf. Piccinini 2008a; Piccinini and Bahar 2013). However, whereas in the case of simple computing systems, such as standard TMs, the individuation strategy seems to be straightforwardly internalist or formal, in the case of more complex (and concrete) computing systems, mechanists defend a *wide functional individuation* scheme.

The idea of a wide functional individuation scheme goes hand in hand with the adoption of an overall mechanistic framework for thinking about computing systems, their functioning, and their applicability to the study of cognition. Recall that, on the mechanistic framework, the function of a complex system is analysed in terms of its component parts, their characteristic functions and organised interaction. Moreover, mechanists claim that in order to *identify* the functions performed by the component parts of a complex system, one must often rely on a set of both top-down and bottom-up considerations. This is because in *identifying* the functions of the component parts of a mechanism one needs to take into account how the complex system itself is embedded in a larger causal mechanism and how its function contributes to yielding the function performed by the higher-order mechanism.

Thus, the mechanistic argument for wide functional individuation seems to be the following. Computational systems are a special type of mechanisms which are *type-individuated* in terms of some of their functional properties. In order to *identify* the functional properties which are relevant for a (complex) system being a particular type of computing system, one needs to appeal to the functions of its component parts and perhaps even to the functions of the larger computing mechanism in which the target system is embedded. The strategy is called 'wide' precisely because in order to establish the functional properties which play a role in the type-individuation of a particular computational system/state one needs to carve them from a wider network of functional relations that hold between the target system, its components, and other (possibly) larger systems.

To further support their proposal, mechanists argue that, 'scientific theories typically individuate the functional properties of mechanisms widely'. Computational theories, as a subclass of scientific theories, are also said to individuate their theoretical posits widely. Otherwise put, in order to distinguish the properties of computing mechanisms that are functionally relevant (both computationally and non-computationally) from those that are not,

> we need to know which of a computing mechanism's properties are relevant to its computational inputs and outputs and how they are relevant. In order to know that, we need to know what the computational inputs and outputs of the mechanism are. That, in turn, requires knowing how the

> mechanism's inputs and outputs interact with their context (cf. Piccinini 2008a, p. 220).

Mechanists also insist that wide functional individuation should be clearly distinguished from individuation based on wide (functional) content, the latter being merely a form of semantic individuation. Along these lines, they write that: '[i]ndividuation based on wide content is one type of wide individuation, but wide individuation is a broader notion. Wide individuation appeals to the relations between a mechanism and its context, relations which may or may not be semantic' (cf. ibid., p. 219).

Thus, mechanists maintain that the correct individuation strategy with respect to computing systems is 'wide individuation that does not appeal to semantic relations' (ibid.). This claim is reinforced by two sorts of considerations. Firstly, mechanists point out that 'the functional properties that are relevant to computational individuation, even when they are wide, are not *very* wide'. They have to do with the normal interaction between a computing mechanism and its *immediate* mechanistic context via its input and output transducers. Secondly, mechanists defend the separation between wide functional individuation and individuation by wide (externalist) contents by claiming that:

> In most of the literature on wide contents, wide contents are largely ascribed by intuition, and theories of content are tested by determining whether they agree with the relevant intuitions. By contrast, under the functional view of computational individuation, the functional properties that are relevant to the computational individuation of a mechanism *are to be found by elaborating mechanistic explanations* under the empirical constraints that are in place within the natural sciences. This establishes the computational identity of a mechanism without appealing to any semantic intuitions (cf. ibid., p. 222).

I submit that neither of the two types of considerations suffices to guarantee that the wide functional view of computational individuation is not yet another version of the semantic individuation hypothesis. Whilst the first motivation simply plays on the ambiguity of the notion of 'wide' content, the second creates a contentious divide between the methodological principles adopted by classical computationalists and those endorsed by mechanists.

That is, the first type of consideration is problematic because it does not actually rule out the possibility that even 'less wide' (functional) properties might count as proper semantic contents. After all, the scale of mental contents proposed by defenders of a semantic individuation strategy need not comprise only the extremes, i.e., (very) broad and narrow contents, but also intermediary values such

as restricted wide functional contents. More importantly, there is an implicit ambiguity about which properties actually count as proper semantic contents on the mechanistic conception. According to mechanists themselves, an external semantics is one which relates internal states of a computational system with entities/properties external to the system itself (cf. Piccinini 2008a). However, on this characterisation, even the 'not very wide' functional properties which are said to be relevant for computational individuation seem to count as semantic properties. Therefore, unless this ambiguity is clarified, wide functional individuation seems to be compatible with a semantic account of computational individuation, *contra* the mechanistic contention.

The second line of defence seems at first blush to be more effective, but it is in fact the source of more trouble for the functional view of computational individuation because it risks confusing the individuation and explanation issues all over again. Appealing to certain methodological principles used by practicing scientists to construct better computational models does not by itself prove that these principles also determine/fix the computational identity of particular physical systems. Still, this observation does suggest a different way of avoiding the problems faced by the wide functional view of computational individuation.

One might argue that in order to *identify* the properties which are relevant from a computational point of view, one needs to take into account a host of factors, including the ways in which the target system is embedded and interacts in certain contexts with other parts of more complex systems. However, this need not conflict with the idea that the properties which effectively determine the *type-identity* of a particular computational system/state are formal or structural properties which hold between the component parts of the system and their internal mode of organisation. In fact, the latter is precisely the view of computational individuation which is directly implied by the notion of generic computation.

Moreover, this latter characterisation supports the contention that there is no substantial disagreement between an internalist view of classical computational individuation and a functional view of mechanistic computational individuation. Recall that, according to an internalist view of computational individuation, the type-identity of a particular computational system is determined exclusively in terms of its formal or structural properties that characterise the way in which the component parts of the system are organised together so that, given a particular type of input, the system is able to generate a specific type of output. The internalist view of computational individuation guarantees that a computational mechanism will keep its computational identity across a wide range of contexts and conditions. In addition, unlike the wide functional view of computational individuation, the internalist account defended in chapter 4 takes a

resolute stance on the separability of the individuation and explanation issues, which in turn allows a better assessment of the principles that govern the application of computational systems/models to the study of different empirical domains.

Thus, I propose that the qualification 'wide' from the wide functional view of computational individuation does, in fact, reflect an additional aspect of the practice of using/constructing computational systems for modelling and theorising purposes, different from mere individuation concerns. Namely, it emphasises the fact that in order to establish or identify the computational system which is most adequate for modelling a particular cognitive phenomenon one needs to appeal to a range of wide functional properties. That is, for the purposes of modelling and *explaining* a cognitive phenomenon, one needs to take into account not only how the system is internally organised but also a series of features which pertain to how the system fits in the wider causal structure of the cognitive task. Some of these features refer to the relations between the computational system and its environment (internal or external) or even to certain semantic features (wide functional contents) whose attribution to certain parts of the computational system would further guarantee its adequacy as a model of a particular cognitive capacity. Rather than fixing the computational identity of a system, these factors seem to play a role is determining whether and when a computational model can *explain* a target cognitive phenomenon.

5.4 THE MECHANISTIC VIEW OF COMPUTATIONAL EXPLANATION

The central idea of the mechanistic view of computational explanation is that computationalist explanations are actually a *sub-species* of mechanistic explanation. My purpose in this section is to examine the reasons offered in support of this thesis. In particular, I am interested in establishing what the mechanistic view of computationalist explanations of cognitive capacities adds to the classical computationalist view of cognitive explanation. With this aim in view, I will adopt the following strategy. First, by analysing the motivations for a mechanistic conception of computationalist explanations, I seek to extract the most salient advantages that seem to recommend the adoption of this proposal. Second, I argue that the case for the mechanistic conception is undermined by two distinct problems. The first derives from equivocations introduced by arguments proposed in support of this view, while the second concerns the consequences that the mechanistic view of computational explanation entails.

5.4.1 *Computational explanations as a class of mechanistic explanations*

Mechanists are keen to point out that one rationale for thinking about computational descriptions and/or explanations within a mechanistic framework is that this strategy guarantees the integration of cognitive computationalist explanations with other explanatory strategies widely employed in other related scientific domains such as physiology and engineering. Mechanists claim that in these scientific fields there is a widespread consensus that the capacities of biological and artificial systems like brains and computers are to be explained mechanistically (cf. Bechtel and Richardson 1993/2010; Craver 2007b).

Thus, proponents of the mechanistic view of computation endorse a broad notion of mechanism, according to which a mechanism is a complex system whose behaviour or function can be decomposed into its component parts, their proper functions, interactions, and organisation. As shown in chapter 3, the corresponding notion of mechanistic explanation is that of a description that elucidates the behaviour of a complex system in terms of a system's components, functions, and organisation. The concise version of the argument for the mechanistic conception of computational explanation claims that since computational systems (digital, analog, neural, etc.) are complex systems made out of rigorously organised parts (i.e., they are particular types of mechanisms), computational explanation is the form taken by mechanistic explanation when the activity of the target mechanism can be accurately described as the processing of adequately typified strings of entities in accordance with the appropriate types of rules.

There is an important assumption implicitly at work in this brief statement of the argument for the mechanistic view of computational explanation. As I pointed out above, mechanists hold that computational explanations should be distinguished from mechanistic explanations *simpliciter*. That is, if computational systems are a special class of physical systems, then the description and explanation schemes appropriate with respect to such systems should also be distinguished from the description and explanation strategies employed in the investigation of non-computing physical mechanisms (e.g., digestive systems, planetary motions, etc.). This commitment is also implicit in the distinction drawn by some mechanists (e.g., Piccinini 2007a) between causal explanations and mechanistic computational explanations:

> Mechanistic explanation, unlike causal explanation simpliciter, distinguishes between a system's successes and its failures. It also distinguishes between the conditions relevant to explaining successes and failures and those that are irrelevant. This gives us the resources to distinguish the properties of a mechanism that are relevant to

> its computational capacities from those that are irrelevant. But *mechanistic structure per se is not enough to distinguish between mechanisms that compute and mechanisms that do not.* [...] The main challenge for the mechanistic account is to specify mechanistic explanations that are relevant to computation (cf. Piccinini 2007a, p. 508).

Sidestepping the potential ambiguity introduced by distinguishing between mechanistic and causal explanation[3], the previous quote emphasises the fact that computational explanations, besides their mechanistic decompositional format, possess an additional feature which distinguishes them from other types of mechanistic explanations. This additional feature is the type of entities or mechanistic components and rules (of interaction) over which genuine computing systems are taken to be appropriately defined.

Although, on various occasions, mechanists seem ready to defer the specification of the computationally relevant mechanistic components to the judgment of the 'pertinent' scientific community, they do sketch a general account that is meant to complete the mechanistic picture of computational explanation. As discussed in section 2, mechanists distinguish between two notions of computation: one abstract and one concrete, and at least three major sub-classes of concrete computation: digital, analog, and neural computations. In each case, they insist that computing systems (abstract and concrete) can be differentiated in terms of the specific type of entities over which computational rules are defined. Thus, what completes the mechanistic picture of computational *explanation* is the idea that a mechanistic decomposition of a system is computational if it specifies the 'right' type of component entities over which the overall functioning of the system is defined.

More specifically, a particular explanation counts as a computational explanation of a digital computer if it describes the functioning of the system and its specific patterns in terms of appropriately defined operations over strings of digits. Similarly, according to mechanists, an explanation of a particular cognitive capacity/process counts as a proper computational explanation if the mechanistic decomposition of the system appeals to appropriately specified neural computations. And so on for the other types of computational systems.

Hence, a direct consequence of this version of the mechanistic view is that a computational model of a particular cognitive process is explanatory only to the extent that it yields a mechanistic decomposition of the process in which the computationally relevant factors are

3 By virtually all mechanistic accounts of explanation, the notion of mechanism provides the key conceptual tool for spelling out the true nature of causal relations, thus providing a more specific framework for articulating causal explanations of various physical phenomena (cf. Machamer, Darden, and Craver 2000; Glennan 1996, 2002, 2010; Craver and Bechtel 2007, etc.)

type-identified as neural computations. More specifically, on this picture, explanatory computational models of cognitive capacities must specify the composition, organisation, and interaction of the variables that are taken to be functionally significant from a computational point of view, viz. the properties of neural spike trains such as firing rate and spike timing. Before assessing this mechanistic explanatory hypothesis in the context of cognitive psychology and neuroscience, I briefly review some of the purported consequences that follow from adopting this mechanistic framework for understanding computational approaches to cognition.

5.4.2 *Four entailments of the mechanistic picture*

Proponents of the mechanistic view of computation claim that their position has a number of attractive consequences which enhance its appeal in comparison with other candidate accounts (e.g., classical and connectionist versions of computationalism). The main entailments of the mechanistic view of computational explanation may be summarised under four distinct labels: objectivity, coverage, genuine explanatory function, and integration.

To begin with, mechanists claim that their account implies that computationalist explanations constitute a distinct type of *objective* scientific explanations. More specifically, they maintain that the mechanistic construal of computational explanations shows both pancomputationalism (Copeland 1996) and the subjectivity of computational descriptions (Putnam 1975; Searle 1992, 2002) to be unacceptable hypotheses. According to mechanists, pancomputationalism trivialises the idea that certain physical systems perform computations, thereby making computationalism about cognitive capacities *a priori* true. Under this view, computationalism about cognitive capacities would just be the trivial application to biological organisms of a general thesis that applies to everything. In response to this potential trivialisation of the computational strategy, mechanists propose that computationalism should be conceived as an empirical hypothesis about the specific types of mechanisms that underlie and explain psychological phenomena.

In other words, mechanists claim that they offer an account on which whether a given mechanism performs a particular computation can be shown to be a matter of fact. This in turn goes against the subjectivity hypothesis according to which computational descriptions are a matter of free interpretation. In brief, the intuition driving the subjectivist reading of the computationalist hypothesis is that any system may be described as performing any computation and there is no further matter of fact as to whether one computational description is more accurate than another. However, the mechanists rightly point out that the modelling practices of computational psychologists,

computer scientists, engineers, and other practitioners are incompatible with the strong relativism of computational descriptions implied by the subjectivist hypothesis.[4] Adequate and potentially explanatory computational descriptions usually have to satisfy a host of constraints and/or norms in any specific modelling context. The claim is that, in all of these fields, there are various ways of ranking the adequacy and potential explanatory value of alternative candidate computational models.

In addition, mechanists claim that their account affords a sharp distinction between genuinely explanatory and non-explanatory computational models of concrete (physical) phenomena. The mechanistic account argues that only mechanistic descriptions which show that the behaviour or function performed by a particular system is the result of the organised interaction of a series of appropriately defined component entities (digits, analog signals, train spikes, quidits, etc.) are good candidates for explanation. This implies that descriptions of complex systems which fail to satisfy either of the two conditions do not qualify as genuine computational explanations *at all*. Neither non-mechanistic descriptions nor mechanistic descriptions which do not mention the appropriate type of component entities qualify as being potentially good computational explanations.

These two mechanistic constraints on computational explanations are supposed to capture the idea that practicing scientists might develop certain computational models without expecting them to have a genuinely explanatory function. Thus, the mechanistic view shows that the mere possibility of developing computational models of particular complex physical systems does not suffice to establish whether the systems in question are proper computing systems or not, i.e., whether the computational models are genuinely explanatory. Mechanists contend that the main problem with adopting a too permissive *modelling view* of computationalism in cognitive science is that it runs the same risk as pancomputationalism does, i.e., trivialising the computationalist hypothesis by making all proposed computational models of cognitive capacities explanatorily relevant. They point out that computational models are used, in cognitive science as in other scientific domains, for various epistemic and experimental purposes, e.g., prediction, confirmation, etc. The double mechanistic constraint is thus meant to block the threat of trivialising the notion of computational explanation by isolating the factors which confer *genuine* explanatory value on a computational model, namely, (i) mechanistic organisation, and (ii) appropriately typified component entities.

As a third major consequence, the mechanistic view of computation is said to accommodate a wide range of computationalist explanatory strategies used in different branches of science. The hypothesis that

4 For a more detailed criticism of the subjectivist reading of computationalism, see e.g., Copeland (1996); Rey (1997); Piccinini (2007a).

computational explanations are mechanistic descriptions which comprise, as functionally significant components, generic computations entails that many types of systems, from abstract Turing machines studied by the theory of computation, to analog, digital, and quantum computers investigated in fields such as computer science and artificial intelligence, and brains studied by cognitive neuroscience, admit of a mechanistic computational description and/or explanation. According to the mechanistic account, the feature that distinguishes these various types of computational descriptions/explanations from one another consists in the type of atomic entities that are responsible for the functioning of the system. Computational explanations of cognitive capacities will characterise the functions performed by complex cognitive systems in terms of the organised interaction of train spikes (i.e., the purported units of neural computation), while the computational explanation of a digital computer will provide a mechanistic decomposition which postulates appropriately defined operations over strings of digits. And so on for the other types of computational systems. Thus, whilst the mechanistic view of computation has a wide coverage and allows one to assess the similarities between mechanistic explanations used in different scientific domains, it also arguably provides a generic criterion for distinguishing between different types of adequate mechanistic explanations.

There is a last important entailment of the mechanistic account of computation. It has been claimed that if computational explanations are a form of mechanistic explanation, then computational explanations of cognitive capacities can be integrated with other types of mechanistic explanations developed at lower levels of neurobiological organisation: cellular, molecular, biochemical, electrophysiological, etc. (cf. Craver and Piccinini 2011). This consequence is taken to refute the classical autonomy thesis according to which higher-level explanations proposed in cognitive psychology are independent from neuroscientific hypotheses and explanations (Fodor 1974; Fodor 1997; Block 1997). Thus, various authors have argued that the mechanistic view of computational explanation corrects the misleading 'two-levelism' hypothesis assumed in most philosophy of mind and psychology (cf. Piccinini and Bahar 2013). Against the two-levelism stance which divides the investigation and explanation of cognitive phenomena into two separable levels of analysis, the cognitive level and the implementational level, mechanists propose that an appropriate approach to psychological phenomena requires a more integrated perspective on the hypotheses proposed at the two purported levels of analysis of cognitive capacities (cf. Craver 2007b; Craver and Piccinini 2011; Piccinini and Bahar 2013). They claim that such an integrated perspective is to be achieved by shifting the focus from cognitive psychology and classical computationalism to cognitive neurobiology and mechanism.

5.5 MECHANISMS VS. COMPUTATIONAL EXPLANATIONS

Although for all the reasons mentioned above, mechanism promises to offer a robust enough framework for analysing the varied landscape of computational approaches to cognition, the view still seems to face important challenges. In what follows, I will focus on two problems which undermine the strong contention that mechanism provides a better model of computational explanation than the classical view. The first problem arises from the fact that, despite their claims to the contrary, mechanists confound the computational individuation and explanation issues. I rehearse the motivations for keeping the two issues apart and trace the consequences of disregarding this distinction for the mechanistic view of computation. In light of these considerations, I reevaluate the relationship between the mechanistic view of computational explanation and the so-called modelling view of computation. The second problem concerns the mechanist's conception of neural computation which is used as a criterion for distinguishing between proper computational explanations and non-explanatory computational models of cognitive and/or neural processes. In relation to this problem, I reopen the issue of the autonomy of abstract (non-mechanistic) accounts of cognitive processes. I conclude, in section 6, by comparing the mechanistic view of computation with the revised version of classical computationalism articulated in chapter 4.

5.5.1 *Individuation versus explanation*

There are two principal motivations for keeping apart the computational individuation and explanation issues. First, the separability thesis allows one to treat the different tools (e.g., computing structures) used in the investigation of cognitive phenomena as independent objects that display a number of interesting properties which support their applicability to the study of different cognitive patterns/phenomena. That is, individuative criteria can and should be viewed as principles that govern the computational systems independently of their application to the study of cognitive phenomena. Second, the separability thesis vindicates the widespread externalist assumptions involved in much cognitive modelling as well as the intuition that the notion of mental content plays an important role in the construction of good models of cognitive capacities. For, on the proposed view of computational explanation, both wide (externalist) and narrow (proximate) mental contents can play a role in the justification of the structures postulated for the explanation of various cognitive patterns.

In addition, by allowing the principles/criteria which determine the type-identity of computational systems postulated in the explana-

tion of cognitive phenomena to be independent of specific semantic assignments, one is in a better position to support the idea that computational explanations function by exhibiting certain stable (more fundamental) features of the cognitive phenomena being investigated. This intuition is also closely related to the idea that an explanatory account should capture as many counterfactual supporting generalisations as possible about the phenomena being investigated. In order to test these counterfactual supporting generalisations one must guarantee that: (i) there are certain features of the system which are stable under a range of well-specified conditions, and (ii) it is possible to specify the varying conditions themselves. On the proposed approach, the first requirement is satisfied by mandating that individuative principles for computational structures take into account only their formal features, whereas the second requirement is met by acknowledging the context-sensitivity of the semantic interpretations assigned to particular computational structures in different modelling contexts. In brief, I claim that the separability hypothesis is necessary in order to make sense of the very idea that explanatory accounts of cognitive phenomena support a variety of *ceteris paribus* generalisations (or laws) concerning the functioning of the cognitive systems under investigation.

I have argued that the mechanistic account of computation can be reinterpreted so as to reflect the separability of the two issues. That is, I have shown that the view of computational individuation which is entailed by the notion of generic computation promoted by mechanists themselves is in line with the internalist account defended in chapter 4. According to an internalist or formal view, the computational identity of a given system depends on a set of formal or structural properties and relations that hold between the component parts of the system and their organisation. In addition, I have argued that the mechanistic reference to 'wide' functional properties is better understood in the context of discussing the applicability and explanatory value of particular computational models of cognitive phenomena rather than in relation to the problem of computational individuation proper. By considering the norms which guide the use of computational systems in the study of cognitive phenomena, one is better placed to vindicate the mechanistic intuition that in order to establish (or justify the choice of) the computational system which best models a given cognitive capacity, one needs to take into account a series of considerations that go beyond the computational identity of a given system.

If I am right that the generic notion of computation supports a formal view of computational individuation, then it seems that the mechanist and the defender of a classical view of computation might be able to agree at least on this point. For the latter need not insist, on this way of carving up the problem, that semantic (wide or nar-

row) contents play a role in computational individuation *per se*, if it is granted that they still are an important part of the explanatory apparatus of computational theories of cognition. After all, the main contention of the classical view is that the semantic or representational level captures a series of interesting cognitive patterns (regularities) which are in need of explanation (Fodor 1980; Pylyshyn 1984). But this sort of commitment need not affect, as the mechanists point out, one's stance on the computational individuation issue. Thus, I contend that both mechanists and defenders of classical computationalism should endorse a purely formal view of computational individuation. In response, the mechanist might still argue that her approach sharpens the correct view of computational individuation and, in addition, provides a series of important insights about how to evaluate the explanatory value of particular computational models of cognitive phenomena. I agree, with the caveat that wide functional contents are better viewed as linking computational models with their target cognitive phenomena rather than fixing the computational type-identity of particular computational systems.

Focusing on the problem of explanation itself, one of the main attractions of the mechanistic account is that it promises to draw a sharp distinction between computational modelling and computational explanation (cf. Piccinini 2008a; Craver and Piccinini 2011; Piccinini and Bahar 2013). According to mechanists, on a *modelling view of computational explanation*, the possibility of constructing a computational model of a given cognitive process is already an index of its potential explanatory power. However, it is easy to show that this claim is problematic. For instance, constructing a computational model requires only that there be a weak input-output equivalence between the model and the target phenomenon, but since input-output equivalences are rather cheap, it follows that one complex system (e.g., a particular cognitive process) admits of a large number of weakly equivalent models. In order to establish which (if any) of these models is a good candidate for explanation, one needs to strengthen the input-output equivalence requirement.

On the mechanistic account of computation, what distinguishes a mere computational model from a proper computational explanation of a particular cognitive capacity is that only the latter, but not the former, provides a mechanistic decomposition of the system into its simpler component parts which in turn can be appropriately characterised as generic computations. Models that fail to specify the componential and functional organisation of a complex system or that are defined over entities which are not of the appropriate type to enter into computational relations do not qualify as genuinely explanatory computational models.

More specifically, according to mechanists, what is special about computational explanations of cognitive capacities is that they postu-

late *neural computations*. Since the latter are taken to be distinct from all other forms of generic computation (cf. section 2.2.3), mechanists conclude that digital, analog, as well as other possible versions of computationalism are inappropriate for studying the nervous system and its functions. Thus, an important consequence of the mechanistic account of computation is that the mathematical theory of computation is not an adequate framework for developing explanatory models of cognitive capacities. Instead, some mechanists suggest that new types of mathematical tools must be developed in order to study the relations holding between the *actual* units of neural computation.

In what follows, I focus on three challenges facing this proposal. I begin by sketching a cautionary argument against the strong reliance of the mechanistic view of computational explanation on certain considerations pertaining to the biological plausibility (or implausibility) of the potential units of neural computation. Then, I question the idea that all computational models which appeal to some notion of neural computation are best understood along mechanistic lines and maintain that there are good reasons to resist imposing any unique framework in order to account for the distinct explanatory contributions of different types of computational models used in different branches of cognitive science and neuroscience. And, finally, I challenge the quick mechanistic rebuttal of the autonomy of higher-order computational models of cognitive capacities. In light of these critical remarks, I reassess the limits of the range of application of the mechanistic framework for understanding computationalist approaches to cognition.

5.5.2 *The limits of biological plausibility*

As shown in the previous sections, the mechanistic account appeals to two distinct factors for establishing the explanatory value of computational theories/models of cognition, viz. the availability of detailed and appropriate mechanistic decompositions of the target physical systems, and the postulation of appropriate computing units, i.e., the entities over which neural computations are appropriately defined. For instance, mechanists like Piccinini and Bahar (2013, p. 475-6) contend that '[g]iven current evidence, the most functionally significant variables for the purpose of understanding the processing of neural signals are properties of neural spike trains such as firing rate and spike timing.' Since these properties are explicitly discussed in the context of their criticism of the hypothesis of digital computationalism, we are led to believe that these authors are committed to the idea that genuinely explanatory neural computations should be defined over these or other similar types of properties of neural spike trains.

However, there seems to be a *prima facie* problem with this mechanistic strategy which ties indiscriminately the evaluation of the explanatory value of all computational models proposed in cognitive science and neuroscience to what biological theory about the organisation of the nervous system currently counts as plausible. Arguments which invoke the biological plausibility of some theoretical posit tend to be misleading because they presume that all potentially new scientific (e.g., biological) hypotheses will have to look plausible in light of our current limited knowledge. But this assumption seems to go against the very notion of scientific discovery and progress.[5] That is, claiming that neural computations must be defined in terms of operations over certain properties of neural spike trains might look like the only plausible option if one takes for granted that future research would not discover other appropriate candidates for being the functional units of neural computation. In order to avoid the fallacy of turning an empirical claim (about what are currently treated as the functional units of neural computation) into a conceptual (a priori) definition of neural computation, one ought to avoid making the notion of computational explanation depend on any particular hypothesis concerning the implementation of the neural processes targeted by particular computational models.[6]

Moreover, the thesis that neural computations should be defined on neural properties such as firing rate and spike timing implicitly assumes that all computational processes (properties) in the biological brain are realised by the same type of structure and in the same way. Whether this is the case or not ought surely to be established through further empirical investigation rather than postulated as part of the definition of neural computation. That is, one should allow for the possibility that different features of neural computing systems (e.g., symbols, rules, structured representations, etc.) might turn out to have distinct implementational bases. Thus, another problem facing the mechanistic argument analysed above is that it disregards several speculative hypotheses according to which certain symbol-like

5 Gallistel and King (2009) proposes a 'cautionary tale' which captures the main moral of this objection. He points out that before the discovery of the structure of DNA (Crick 1953), the gene was a deep biochemical puzzle and that its 'reality' was doubted by a number of biochemists. This was because: 'A gene was assumed to have two utterly mysterious properties: it could make a copy of itself and it could direct [...] the synthesis of other molecules. Because the properties of a gene assumed by geneticists made no biochemical sense, some biochemists simply refused to believe in its physical reality, despite what I think almost anyone in retrospect would argue was a very large body of evidence and theory arguing that there had to be such a thing' (Gallistel and King 2009, p. 281).

6 The same sort of considerations apply to arguments from biological implausibility (the flip side of biological plausibility). Recall that part of the mechanistic argument against classical computationalism consists in noting that we do not yet know how to implement digital computation in the neural tissue of the brain. But I contend that such claims are just misleading appeals to ignorance that have no demonstrative force.

features of neural computation might have a molecular (or even sub-molecular) implementational basis (e.g., Gallistel and King 2009; Gallistel and Matzel 2013; Marcus 2009, 2013). For instance, Gallistel and King (2009, p. 280) suggest that plausible mechanisms for memory might be found in certain 'ingenuous adaptation[s] of the molecular machinery that is already known to have an information-carrying function'. They speculate that at the sub-molecular level, changes in nucleotide sequences in either DNA or RNA might constitute a possible mechanism for memory, whereas at the molecular level, such a mechanism might be realised by a rhodopsin-like switch molecule.[7]

Although both of these speculative claims would require much more theoretical and experimental work before one could properly assess their empirical adequacy, they still serve to illustrate two important points about the mechanistic arguments from biological plausibility. Firstly, the mechanists ought to be careful not to rule out the possibility that other levels of organisation and/or resolution than the level of neural spike trains and their properties might be relevant for the application of different computational notions to the study of cognitive and neural processes. Secondly, the previous remarks caution against the impulse to ground (almost exclusively) the explanatory value of all computational models constructed in different branches of cognitive science and neuroscience in certain (potentially defeasible) hypotheses about the plausible biological basis of neural computation.[8] Thus, my contention is that appeals to biological plausibility by themselves cannot guarantee the explanatory value of a given computational model of a particular cognitive or neural process.

However, there are at least two promising response strategies that mechanists might use in order to avoid this sort of criticism. On the one hand, mechanists might retort that their account of computational explanation in the cognitive and neuroscientific domain does not commit them to any specific hypothesis about the realisation basis of neural computations. Along these lines, they write that: '[i]t

7 Whilst highly speculative, both realisation hypotheses avoid the second major problem faced by the mechanistic conception of neural computation. Namely that of accounting for the difference between the speed of neural computation and that of nervous signal transmission (which seems to travel several orders of magnitude slower). Modelling neural computation at the cellular or circuit-structure level necessarily raises the problem of accounting for the differences between the two temporal profiles. However, if the (computational) functions and structures that are currently modelled at the level of cellular or circuit structure can in fact be implemented at the level of molecular structure, then this would deal with what is presently a serious limitation of neuroscientific models. And this is because in the case of the molecular (and sub-molecular models) much less time and space would be wasted transmitting signals over long distances (as in the cellular models).

8 As Gallistel and King (2009, p. 287) somewhat strongly point out: '[u]ntil the day comes when neuroscientists are able to specify the neurobiological machinery that performs this key [memory] function, all talk about *how* brains compute is premature.'

does not follow that we should consider only algorithms that can be implemented using spiking neurons and abandon immediately any other research program.' Instead, they claim that their position entails that: 'anyone seriously interested in explaining cognition should strive to show how the computations she or he postulates may be carried out by neural processes, to the extent that this can be made plausible on current neuroscience. The better an explanation of cognition is grounded in neural computation, the better the explanation' (Piccinini and Bahar 2013, p. 480).

Whilst the first part of the previous quote seems to allow for the possibility of developing relatively autonomous computational models of cognitive and/or neural processes, the second part reinstantiates the dependency between the explanatory value of particular computational models and specific hypotheses about the neural mechanisms underlying the phenomena/patterns targeted by these models. In other words, mechanists seem to be committed to the idea that all computational explanations should be constrained by mechanistic norms. Moreover, mechanists also seem to imply that explanatory prowess is always gained by specifying more details about the mechanisms underlying the computational characterisations of certain cognitive and/or neural processes. In the following section, I will challenge the idea that all computational models used in cognitive neuroscience should conform to mechanistic strictures in order to count as genuinely explanatory accounts. In its place I propose a moderate version of the autonomy of computational explanations used in the cognitive and neuroscientific domains, respectively.

Still, mechanists can appeal to a more reasonable strategy in defense of their position. Namely they could maintain that particular neurobiological hypotheses are ontic norms which guide the construction of certain kinds of computational models of cognitive or neural processes. Thus, rather than claiming that all computational models ought to conform to mechanistic strictures, a more robust position would be that biological plausibility considerations play a partial role in the evaluation of the explanatory value of particular computational models. That is, there are contexts in which it is possible to *coordinate* particular computational hypotheses about certain cognitive or neural processes with information pertaining to the neurobiological machinery/mechanisms supporting them. This strategy can yield, as mechanists point out, potentially explanatory models of the capacities or processes under investigation.

However, the desideratum of achieving this sort of coordination does not by itself imply that only models/theories which can be thus coordinated can be genuinely explanatory. Rather, the limited (or partial) applicability of neurobiological constraints on computational modelling is consistent with the idea that computational models might play an explanatory role independently of any mechanistic

constraints. I claim that this proposal avoids the problems of making the notion of neural computation depend on any specific empirical neurobiological hypothesis concerning the realisers of particular types of neural computations.

Otherwise put, granting that only certain computational models of neural processes are constrained by mechanistic norms yields a notion of neural computation which, whilst distinct from that of digital or analog computation, is not tied to any particular realisation hypothesis about the mechanisms underlying these computations. That is, I claim that one is better off taking the notion of neural computation to stand for any formally describable input-output transformation that neural systems are capable of performing. Note that whilst this way of characterising the notion of neural computation is compatible with the broad outlines of the mechanistic strategy of distinguishing between different types of computing notions, it also acknowledges the open-textured nature of this notion, whose theoretical definition should still be considered as work in progress (cf. Chirimuuta 2014).

5.5.3 *The case of canonical neural computations*

In a recent paper, Chirimuuta (2014) has argued that practicing computational neuroscientists make use of a distinctive (non-mechanistic) explanatory style, viz. *efficient coding explanation* in order to account for certain salient properties of neural processes. Against mechanists like Kaplan (2011), Piccinini (2008a), and Craver and Piccinini (2011), she claims that computational explanations in neuroscience are better understood as *minimal interpretative* models that are not necessarily constrained by the mechanistic norms of explanation. According to Chirimuuta (2014), intepretative models ellucidate *why* a particular neuronal type or brain area is organised in a certain way by appealing to efficient coding principles. She maintains that whilst computational principles are central to this style of modelling and explanation, detailed mechanistic descriptions are not required and can even impede the success of explaining certain properties or patterns of neural processing. The resulting explanatory models are said to be minimal because in order to design and test them, computational modellers usually abstract away from most of the biophysical details of the target neural systems (cf. Chirimuuta (2014); see also Batterman 2000, 2002).

At this point, it is worth stressing that the *interpretative minimal* view of computational explanation defended by Chirimuuta (2014) seems to appeal to a more robust, theory-neutral, notion of neural computation than the one proposed by defenders of the mechanistic position. Thus, if her proposal is on the right track, it supports the idea that the notion of neural computation can play a role in

constructing explanatory models/theories in neuroscience independently of any mechanistic assumptions. Chirimuuta (2014) illustrates her view of computational explanation with the help of two important case studies: (i) the contrast normalisation model of the visual primary cortex and (ii) the Gabor-model of V1 receptive fields (RFs). Against Kaplan (2011) who maintains that these and similar models offer 'mere' phenomenal descriptions of the target neural properties or patterns, she argues that these models exhibit a distinct (non-mechanistic) style of explanation. In what follows, I will focus only on the case of normalisation models, but very similar considerations apply to her other case studies as well.

As her starting point, Chirimuuta (2014) notes that a large body of literature addressing the methodological and explanatory concerns of computational neuroscience emphasises the importance of abstraction and idealisation for the purposes of modelling and explaining certain salient neural properties and/or patterns (e.g., Sejnowski et al. 1988; Trappenberg 2010; Steratt et al. 2011). That is, an important part of the community of computational neuroscientists seems to favour the hypothesis that at least in certain contexts, minimal models can provide better explanations of certain salient features of the complex neural systems being investigated.[9] One case study which seems to support this hypothesis is the *contrast normalisation model* of the primary visual cortex. Presented by David Heeger in 1992, the normalisation model provides a quantitative account of the response properties of simple cells in the primary visual cortex. According to this model, 'each simple cell has linear excitatory input originating from the LGN, and in addition it receives inhibitory input from nearby neurons in the visual cortex' (cf. Chirimuuta 2014, p. 136). Although *prima facie* this model seems to be a mechanistic sketch which provides an incomplete description of some of the components of the mechanism that supports the response profiles of simple cells in the primary visual cortex, Chirimuuta (2014) points out that the normalisation model is currently treated as an instance of *canonical neural computations* (CNC) performed by the brain. The latter are typically conceived of as 'standard computational models that apply the same fundamental operations in a variety of contexts' (cf. Carandini and Heeger 2012, p. 51). Thus, the normalisation model together with other CNCs (e.g., linear filtering, recurrent amplification, exponentiation) is taken to capture a computational operation which is applied

9 In discussing the notion of 'minimal' models, Chirimuuta (2014) is careful to distinguish between what she calls *B-minimal models* (cf. Batterman 2002) and *A-minimal models* (cf. Weisberg 2007; Strevens 2004, 2007). The latter are typically treated as providing a kind of causal/mechanistic explanation, whereas the former are taken to make their explanatory contributions to scientific investigation independently of mechanistic considerations (cf. Batterman 2002a, 2002b). (For a more detailed discussion of the distinction between these two classes of minimal models, see Chirimuuta 2014, pp. 141-147.) For the purposes of this section I will focus exclusively on the notion of *B-minimal model*.

by the brain in different anatomical regions (corresponding to different sensory modalities) and which can be subsequently described independently of any assumption about its potential biophysical implementations.

Drawing on the physiological and behavioural evidence cited by Carandini and Heeger (2012) that shows that normalisation models can be successfully applied to a wide variety of modalities, brain regions, and species, Chirimuuta (2014) claims that the explanatory value of such models does not depend on the specification of the particulars of the neural mechanisms underlying these types of computations. Instead, she argues that their distinct explanatory contribution is better understood as consisting in the identification and characterisation (at a higher level of abstraction from mechanistic considerations) of certain stable properties or patterns of neural processing observed across different sensory modalities and brain regions. As such, these models support the formulation and testing of a very wide class of counterfactual generalisations concerning the functioning of various neural systems. This observation helps locate the *interpretative minimal model of computational explanation* defended by Chirimuuta (2014) in the broader *difference-making* account of explanation proposed by Woodward (2003, 2013).

I conclude this section with two final considerations about the model of computational explanation sketched above. First, the idea that canonical neural computations can play the role of explanatory structures in the context of cognitive neuroscience independently of specific mechanistic hypotheses does not imply that 'CNCs are completely independent of biophysical implementation or that CNCs can best be studied in full isolation from mechanistic considerations' (Chirimuuta 2014, p. 139). Rather, the central claim of the *interpretative minimal* view is that computational modelling makes available a *distinct style of explanation* that can be successfully used in cognitive neuroscience. Moreover, this view of computational explanation does not imply that minimal computational models should be regarded as complete accounts of cognitive or neural processing. Their explanatory partiality is in principle consistent with the mechanistic desideratum of coordinating (or integrating) different types of explanatory models which target various aspects of the same cognitive and/or neural processes. However, as pointed out above, this desideratum should not be taken to restrict the explanatory power of computational models to those for which one has appropriate linking-ready mechanistic hypotheses.

There are at least two good reasons to resist imposing mechanistic norms on all computational models used in cognitive neuroscience. On the one hand, since such linking-ready hypotheses are not always available, the mechanistic position risks imposing too strong of a constraint on the class of explanatory models currently used in cognitive

neuroscience. On the other hand, as suggested above, the explanatory contributions of at least certain computational models do not depend on the availability of mechanistic descriptions at all, but rather on their ability to capture and account for certain stable general features of neural processing.

Second, a number of researchers have proposed that one important role of canonical neural computations is to provide 'important simplifying insights into the relationship between neural computations and behaviour' (Angelaki et al. 2009). For instance, it has been suggested that a wide variety of observed attentional modulation effects can be explained in terms of a unique computational model of contrast gain control (ibid.). More generally, the idea is that due to their abstract (non-mechanistic) character, canonical neural computations, such as normalisation models of different neural systems create a bridge between hypotheses about the behavioural patterns targeted by cognitive research and more detailed mechanistic models of neural computation. In the following section, I consider another class of computational models used in cognitive science that are grounded in classical computationalist assumptions and which make very similar bridging claims. By focusing on this class of computational models, I aim to formulate a more reasonable thesis concerning the autonomy of abstract computational models of cognitive capacities than the 'two-levelism' hypothesis typically targeted by mechanistic arguments (cf. Piccinini and Craver 2011; Piccinini and Bahar 2013).

5.5.4 *The relative autonomy of computational models of cognitive capacities*

The challenge of bridging the gap between behaviour and the neurobiological machinery that underpins it is more serious than cognitive neuroscientists are sometimes willing to admit. As an illustration, consider the fact that even if the orthodoxy in current neuroscience is that the mechanisms of long term potentiation (LTP) constitute the basis of memory, this hypothesis by itself cannot account for any of the various cognitive behaviours which seem to require the exercise of different kinds of memory systems. That is partly the reason why researchers need to appeal to a richer set of conceptual and experimental tools in order to be able to explain different complex cognitive behaviours. For instance, a more tractable approach to the various problems raised in the study of memory seems to be provided by the classical computationalist framework. Gallistel and colleagues (e.g., Gallistel 1993; Gallistel and King 2009; Gallistel and Matzel 2013; Gallistel and Balsam 2014) have argued that positing an addressable read-write memory structure would help explain a host of behavioural patterns observed in study of animal spatial learning and navigation. The notion of a read-write memory which is used as an explanatory

structure in the computational models proposed by these researchers mirrors the type of read/write (fetch/store) memory encountered in virtually all engineered (digital) computing machines. In a digital computer, this type of memory is a universal device which allows information acquired at different times to be stored and used to influence the current behaviour of the machine. That is, the results of the computations performed on earlier inputs to the machine are stored in memory (written), so that when they are needed in future computations they can be retrieved from memory. In what follows, I argue that appeals to the digital notion of a read-write memory facilitate the explanation of a host of cognitive behaviours and that the resulting explanatory strategy is consistent with the general aims and concerns of mechanistic modelling in cognitive neuroscience.

Since analysing the details of these computational models lies beyond the scope of this paper, I will support my claim by focusing on the case of insect navigation.[10] Most insects have a home base from which they leave in search for food and to which they eventually return. This type of behaviour involves navigation which in turn, it has been argued, relies on the storage of information acquired from experience. For instance, a large number of experiments show that if an ant who sets on a particular outward foraging path were captured and displaced into an unfamiliar territory, it would still run the same compass course it ran in returning to its nest for approximately the same distance and then begin the search for its nest. It has been argued that the ability to run a prescribed course for a prescribed distance through an unfamiliar territory implies dead reckoning which is understood as the integration of the velocity vector with respect to time to obtain the position vector as a function of time. In discrete terms, 'dead reckoning requires the summation of successive displacement vectors, with the current displacement vector being continually added to the sum of the previous displacement vectors', which can be further understood as an instance of the composition of functions. This observation is important because it supports the idea that navigation (viz., dead reckoning which is the recursive composition of the simple function of addition) 'requires a memory mechanism capable of carrying forward in time the information acquired from earlier experience (earlier displacements) in a form that permits that information to be integrated with subsequently acquired information (later displacements)' (cf. Gallistel 2012, p.46).

More generally, Gallistel et al. have forcefully argued that in order to explain the various behavioural patterns observed in the wide variety of experiments on insect navigation one needs to postulate that these organisms are able to learn representations of spatial locations and directions, and various time durations and time intervals. This

10 For a more comprehensive review of the behavioural literature supporting this style of computational modelling, see Gallistel and Matzel 2013.

theoretical outlook on modeling spatial learning and navigation further implies that the brain has one or more spatial coordinate systems that encode locations in one or more frames of reference, and also that it can perform distance and direction estimating computations (cf. Gallistel 1993; Gallistel 2012; Gallistel and Matzel 2013; Gallistel and Balsam 2014). Moreover, since dead reckoning (or path integration) and piloting are thought to be essential for different kinds of spatial navigation, these computational models postulate that the brain must be able to implement (perhaps at the cellular or molecular level) these abstract operations. Thus, classical computationalist models/theories which posit abstract computational structures and operations in order to account for the behavioural patterns observed in insect navigation is not entirely cut off from mechanistic concerns. In fact, proponents of such classical computationalist models often insist on the importance of developing neurobiological models that could be coordinated with higher-level (abstract) models of particular cognitive capacities. However, given the many challenges facing the search for such linking-ready models, these researchers point out the need to appeal to a distinct explanatory strategy for dealing with the complexity of cognitive behaviours exhibited by various living organisms.

Thus, rather than having to maintain that explanations of particular cognitive capacities are something that can be achieved only in the horizon of a future complete neuroscience, an appropriate view of the explanatory aims and purposes of cognitive and neuroscientific research should acknowledge the distinct character of computational models developed in the various branches of cognitive science. I claim that in order to avoid a paralysing scepticism concerning the prospects of cognitive neuroscientific research one should endorse a moderate thesis concerning the explanatory autonomy of the models developed at higher levels of abstraction than mechanistic descriptions.

This form of explanatory autonomy would not be grounded in the mistaken assumption that higher-order models ought not be constrained by any sort of hypothesis concerning the biophysical bases of cognitive processes, but rather in the acknowledgment that abstract computational models play an explanatory role by reducing the gap between behaviour and its underlying neural mechanisms. For instance, classical computational models postulate a series of abstract structures and operations which capture certain stable patterns or properties exhibited by functioning cognitive organisms. Otherwise put, by appealing to certain abstract computational structures one is a position to test and confirm a wide variety of regularities that seem to govern the behaviours of many cognitive systems. Furthermore, as mechanists themselves are keen to point out, these abstract models can help design or refine particular mechanistic descriptions of

the biophysical underpinnings of the cognitive capacities exhibited by different living organisms.

However, as argued above, one should not take this aim of cognitive research to constitute the sole valid criterion for evaluating the explanatory value of models of cognitive or neural processes. Classical computational models used in cognitive research differ from other modelling approaches in that they postulate certain abstract structures and operations (representations and rules) which help express a series of important regularities governing the cognitive behaviour of living organisms. In fact, as shown in chapter 4, classical computationalists (e.g., Pylyshyn 1984; Gallistel 1993) have typically insisted that the main motivation for postulating such explanatory structures is precisely because they allow the formulation of a host of relevant counterfactual supporting generalisations about the cognitive behaviours being investigated. I submit that in so far as classical computational models can be viewed as providing a bridge between the analysis of the complex cognitive behaviours of living organisms and more detailed causal/mechanistic descriptions of the biophysical level of organisation of these organisms, they earn their explanatory keep. Moreover their relative autonomy from mechanistic norms is not the expression of the sort of dogmatic 'two-levelism' criticised by mechanists but rather a necessary condition for the explanatory productivity of the field of cognitive science as a whole.

5.6 CLASSICAL COMPUTATIONALISM, MECHANISM OR BOTH?

Proponents of the mechanistic view have argued that their account of computational individuation and explanation provides a better alternative to classical computationalism on at least three grounds. Firstly, they claim that the mechanistic account allows for a larger variety of computational architectures that may be used in the investigation of the structure of particular cognitive capacities and neural processes. Secondly, mechanists hold that the functional view of computational individuation avoids all the problems facing the classical computational individuation strategy. And, thirdly, the mechanistic account is said to provide a more accurate characterisation of the explanatory strategies used in different sub-branches of cognitive science. Despite the strong revisionary tone of these claims, I contend that the mechanistic account has more features in common with classical computationalism than is generally acknowledged. Also, I have argued that some computational models of cognitive and/or neural processes can be said to have an explanatory value despite the fact that they are not directly constrained by mechanistic norms.

One of features that mechanistic and classical computationalism have in common is that they both seem to be committed to a non-semantic, formal view of computational individuation. According to

this view, the type-identity of particular computational systems depends solely on certain formal or structural properties of the systems in question. In particular, their computational type-identity does not depend on the semantic contents that their internal states might be taken to possess as well. I have argued that this view of computational individuation can be easily derived from the generic notion of computation proposed on the mechanistic account and that it also reflects the classical computationalist commitment to the formality constraint thesis (cf. Fodor 1980). According to the latter, computational processing is sensitive only to the formal (syntactic) or structural properties of certain physical systems.

Secondly, so far as the explanatory strategies entailed by each of the two views of computation are concerned, they are far from being incompatible. In fact, the classical computationalist strategy provides a series of insights into the modelling and explanatory practices of cognitive scientists that are consistent with the mechanistic picture of computational explanation (Craver and Piccinini 2011). That is, I have insisted that most researchers working in computational cognitive science and neuroscience seek to develop linking-ready computational models/theories that could be coordinated with mechanistic hypotheses/theories concerning the biophysical bases of cognition. However, against the mechanistic contention that all computational explanations used in cognitive neuroscience are a sub-species of mechanistic explanations, I have argued that computational modelling affords a distinct explanatory style which is relatively autonomous from mechanistic norms.

In support of this contention, I have focused on two case studies in which the notion of computation is used to construct explanatory models of neural and/or cognitive processes. First, the example of canonical neural computations (Carandini and Heeger 2012; Chirimuuta 2014) illustrates how the notion of neural computation can figure in abstract explanations of certain general patterns or properties of neural systems independently of any specific assumption about the biophysical implementation of neural computation. Thus, following the discussion in Chirimuuta (2014), I have pointed out that canonical neural computations like the contrast normalisation model are best understood as minimal (non-mechanistic) models which explain why a large variety of neural systems exhibit certain salient properties or patterns. Second, the example of classical (symbolic) computational models of insect navigation (Gallistel 1993; Gallistel and King 2009) shows the productivity of postulating certain abstract computational structures and operations in order to account for the behavioural patterns observed in insect navigation. In connection with the latter type of example, I have argued that the relative autonomy of computational models of cognitive capacities developed at higher levels of abstraction from their biophysical implementation is a necessary con-

dition for reducing the significant gap between behaviour and neural mechanisms that cognitive neuroscience is currently facing.

Therefore, I claim that although classical computationalism uses a top-down methodological strategy in order to model and explain different cognitive capacities, it is not guilty of ignoring altogether the constraints imposed from a bottom-up, neurobiologically inspired perspective. A better way to understand the commitments and aims of classical computationalism is to view it as a modelling and explanatory strategy which seeks to create a link between the description of complex cognitive behaviours or capacities and the mechanistic characterisation of their neurobiological/biophysical underpinnings. At different stages in this chapter, I have pointed out that mechanists tend to downplay the difficulty of bridging the gap between behaviour and neural mechanisms by claiming that neural computations provide the building blocks for explanations of both cognitive and neural processes. The main problem with this tendency is that it distorts both the successes and challenges facing the field of cognitive neuroscience. A more fruitful perspective, I have argued, should acknowlege that practicing cognitive scientists and neuroscientists appeal to a variety of modelling strategies, each of which yields partial explanatory accounts of the phenomena or systems being investigated. In particular, computational cognitive science and neuroscience often appeal to distinct forms of computational explanation which are not directly constrained by mechanistic norms. That is not to deny the mechanistic concerns of these fields, but rather to recognise the productivity of using different explanatory styles each of which reflects the limits on the range of application of any particular explanatory strategy or structure. In light of these arguments, I conclude that not all types of computational explanations used in the cognitive domain are best understood as a sub-species of mechanistic explanation.

Lastly, it should be noted that despite the strong stance that mechanists take against digital computationalism, they do in fact preserve the central distinction defended by classicists in their broader notion of generic computation. More specifically, the mechanistic definition of generic computation presupposes that one is able to distinguish between a specific kind of entities (i.e., symbols) and operations (i.e., general rules) defined over appropriately typified entities. This in turn supports the idea that there is a substantial element of continuity between the digital computationalist hypothesis and the neural computationalist one. In consequence, I claim that, at a general level, the notion of neural computation constitutes a less radical departure from the classical notion of computation than usually argued.

In fact, the distinctive contribution of the mechanistic view of computational explanation seems to be that it identifies a different type of structure, viz. neural computations, which may be used to explain

certain aspects of the neurobiological mechanisms underlying cognitive processing. Moreover, mechanists make a compelling case that in order to evaluate the explanatory value of particular theories/models which deploy such a notion, one must take into account a series of hypotheses concerning the organisation of the nervous system at the systems and cellular levels. These hypotheses and/or principles function as local norms which constrain and direct the construction of adequate and potentially explanatory models of cognitive phenomena. In the following chapter, I propose to analyse yet another style of computational approach to cognition, which shares various elements with both classical computationalism and mechanism.

6

CONNECTIONIST APPROACHES TO COGNITION

6.1 INTRODUCTION

The previous two chapters have focused on two major philosophical accounts of computationalism, analysing their underlying theoretical principles and associated hypotheses concerning the nature and structure of cognitive explanation. In order to arrive at a more comprehensive picture of the computational tools and techniques currently used within cognitive science, this chapter will focus on the theoretical principles and practical implications of *connectionist* approaches to cognition.

In the last thirty years, connectionist models, also known as artificial neural networks (ANN) or parallel distributed processing (PDP) models, have been applied to a diverse range of cognitive capacities, including memory, attention, perception, action, language, concept formation, and reasoning (e.g., Rumelhart, McClelland, and PDP Research Group 1986; Elman 1996; Christiansen and Chater 2002). Although many connectionist models target adult cognitive processes, connectionism has been promoted primarily for yielding novel and important insights into the nature of learning, thus urging an increased focus on acquisition and developmental phenomena. Supporters of this computationalist framework have argued that connectionist modelling provides an alternative approach to the study of cognition which differs in crucial respects from that afforded by classical computationalism.

Connectionists reject the assumption that the internal computations underlying cognitive processing must be characterised in terms of complex symbolic structures and rules defined over appropriately typified symbols. Whilst a separate issue, connectionism is also often taken to contest the classical innateness hypothesis according to which an important part of the machinery underlying cognitive processing is innate (e.g., Chomsky 1975; Laurence and Margolis 2001; Berwick et al. 2011). In contrast, they argue that: (i) all cognitive phenomena arise from the propagation of activation among simple neurone-like processing units and (ii) such propagation is mediated by weighted synapse-like connections between units. In light of these two central tenets, connectionists claim that one important reason for

preferring the adoption of a connectionist (over a classical) framework is that connectionist models are neurologically more realistic (plausible) than symbolic computational models of cognitive capacities.

The arguments developed in this chapter pursue three interrelated goals: (1) to clarify the relationship between connectionist and classical (symbolic) approaches to cognition, (2) to assess the explanatory roles of connectionist notions and principles in modelling particular aspects of cognitive processing, and (3) to draw out the main challenges facing the exclusivist adoption of a connectionist framework within cognitive science. I take the last of these to be a strong presupposition implicit in both past and current philosophical discussions of connectionist approaches to cognition. I will argue that this claim is in part the result of a strong polarisation between classical and connectionist versions of computationalism. By mitigating this dichotomy, I seek to shed light on both the similarities and differences between the two types of strategies of explaining cognitive phenomena.

The structure of this chapter divides into three distinct parts. In section 2, I begin by surveying the main concepts and principles that constitute the core of the connectionist framework. As in the previous two chapters, I adopt an argument strategy which distinguishes questions that pertain directly to the theoretical framework of connectionism from questions regarding the applicability of connectionist concepts and tools to the study of cognition. I will explore first the individuation problem, seeking to draw out the relevant distinctions between classical computational individuation and connectionist individuation. Next, I discuss the representational problem, which pertains more directly to the issue of the applicability of connectionism to the study of cognition. This will lead, in the last part of the section, to a preliminary discussion of some of the signature features of the explanatory strategy promoted by connectionists in the investigation of cognitive phenomena.

Having discussed the theoretical underpinnings of connectionism, in section 3, I adopt a practice-based perspective in order to analyse in more detail the strengths and limitations of connectionist models of cognitive capacities. First, I briefly survey some of the most important connectionist theses that are standardly taken to challenge the adequacy of classical computational models/theories of cognition. I then illustrate these claims via two influential classes of connectionist models developed and refined in the cognitive literature, viz. connectionist models of English-past tense inflection (Rumelhart, McClelland, and PDP Research Group 1986), and word recognition (Elman 1996). The critical analysis of these models serves three purposes: (i) to show how connectionist methods are actually applied in the investigation of specific aspects of cognition, (ii) to illustrate some of the most polemical connectionist claims made against symbolic computational modelling, and (iii) to pin down the main challenges currently

faced by connectionist accounts of cognitive phenomena. In section 4, I return explicitly to the problem of connectionist explanation in order to sketch a more adequate way of thinking about the contributions made by connectionist modelling to the study of cognition.

6.2 THE MAIN TENETS OF CONNECTIONISM

This section explores the central notions and principles underlying connectionist approaches to cognition. Connectionism claims that the internal computations which underlie cognitive processing are carried out by a set of simple processing units that operate in parallel and affect each other's activation states via a network of weighted connections. In order properly to evaluate the ambitious claim that connectionism can provide the foundation for a new paradigm of computational theories of cognition that supplants the classical (symbolic) framework, I propose to elucidate first the central theoretical tenets of the connectionist framework. After introducing the principal design features of a connectionist architecture (section 2.1), I discuss the individuation issue in a connectionist setting (section 2.2). More precisely, I survey a number of potential strategies that might be used for the type-individuation of particular connectionist networks. Via this analysis, I seek to pin down both the relevant differences as well as the similarities between the individuation of connectionist systems, on the one hand, and classical computational systems, on the other.

I then proceed to tackle the representationalist problem as it pertains to the applicability of connectionist architectures in the study of cognition (section 2.3). There are two primary motivations for a critical analysis dedicated to this issue. First, the postulation of representational contents seems to play an important 'bridging' role in the application of connectionist principles to the study of particular cognitive problems. Second, the representationalist character of connectionist models/theories has been taken by many philosophers to constitute the litmus test for the adequacy of connectionism as a proper cognitive-level theory. Thus, my aim is to clarify the roles played by representational contents in the connectionist framework and how they constrain proposed connectionist explanations of particular cognitive phenomena. The preliminary conclusions concerning the nature and structure of connectionist explanations of cognition (section 2.4) will be complemented, in section 3, by a critical analysis of some paradigmatic connectionist models of higher-order cognitive processes.

6.2.1 *Basic features of connectionist networks*

Current connectionist architectures applied in the study of cognition span a wide range of systems which nevertheless share various prop-

erties and types of components. In what follows, I provide a rough survey of some of the most salient principles and properties of connectionist networks. The principal purpose of this section is to identify the general (theoretical) features which are most likely to play a role in the type-individuation of different connectionist systems.

Following the classical exposition of the connectionist framework (cf. Rumelhart, McClelland, and PDP Research Group 1986; Bechtel and Abrahamsen 1991), I discuss next seven prominent features of connectionist or artificial neural networks (henceforth, ANNs). First, all ANNs comprise a set of *processing units* (u_i) which are often distinguished into input, output, and hidden units. Units are standardly organised into levels or layers which are connected in various ways to one another, so that the processing activity in one layer affects the activity of some other layers of the connectionist network. Second, a network of such processing units is characterised at any given time (t) by an *activation state* (a). More specifically, the state of a set of processing units will be represented by a vector of real numbers $a(t)$. A typical assumption about this representation format is that the activation level of simple processing units varies continuously between 0 and 1.

The notion of activation level is strongly related to a third feature of connectionist networks, namely that of *pattern of connectivity*. The activation state of one unit can be affected by the activation state of another unit in function of the value of the connection strength between the two units. Most commonly, the strength of the connections between two units (i and j) is represented by a matrix W of weight values (w_{ij}). If a particular network comprises more than one type of connection, then the patterns of connectivity may be represented by multiple matrices. For instance, one matrix may specify excitatory connections between units and another may specify inhibitory connections. In principle, the weight matrix allows that every processing unit in the network may be connected to every other unit in the network. However, as hinted above, units are typically arranged into layers (e.g., the input, hidden, and output layers) which in turn may be fully or partially connected to one another.[1]

The rule which specifies how activation states are propagated in a network constitutes the fourth prominent feature of any ANN. Such a rule takes the vector $a(t)$ of output values for the processing units sending activation and combines it with the connectivity matrix W to produce a net input into each receiving unit. The net input to a receiving input is given by the following formula: $net_i = W \times a(t) = \sum_i w_{ij} a_j$. This rule is used to determine how the net inputs to a given unit are combined to produce its new activation state. The fifth,

1 For instance, in a three-layer feedforward architecture, i.e., where activation passes in a single direction from input to output, the input layer would be fully connected to the hidden layer which in turn would be fully connected to the output layer.

and closely related, feature is the activation function F which, as just suggested, derives the new activation state, i.e., $a_i(t+1) = F(net_i(t))$. A key property of the activation function is that it can vary, i.e., it can be either a threshold function (so that the unit becomes active only if the net input exceeds a certain given value), linear, Gaussian, or a sigmoid function, depending on the particular type of artificial network one seeks to build.[2]

The sixth key feature of an ANN is the so-called *learning algorithm* (or function) that modifies the patterns of connectivity as a function of experience. Almost all learning algorithms defined within connectionism can be shown to be variants of the Hebbian algorithm (cf. Hebb 1949). The driving idea behind this very simple learning function is that the weight between two units must be altered in proportion to the units' correlated activity. That is, if a unit u_i receives input from another unit u_j, and both are highly active, then the weight w_{ij} from u_i to u_j is strengthened, following the simple rule: $\triangle w_{ij} = \eta a_i a_j$, where η is a proportionality constant also known as the *learning rate*. This simple rule is slightly modified if an external target activation $t_i(t)$ is available for a unit i at a time t to reflect the disparity between u_i's current activation state $a_i(t)$ and its target (i.e., desired) activation state $t_i(t)$, thus yielding the so-called *delta rule*: $\triangle w_{ij} = \eta (t_i(t) - a_i(t))a_j$.

However, if the network includes hidden units, one cannot determine target activation states directly, so the weights of such units may be modified solely by certain variants of the Hebbian learning algorithm or by the backpropagation of error signals from the output layer. Backpropagation allows one to determine, for each connection weight in the network, the effect of changing its value for the overall network error. The general rule for changing the strengths of connections consists simply in adjusting each weight in the direction that would tend to reduce the error, and change it by an amount proportional to the size of the effect that the adjustment would have. In the case in which the network comprises multiple layers of hidden units, this process can be iterated as many times as required. First, error derivatives are computed for the hidden layers closest to the output layer; from these, derivatives are computed for the next deepest layer, and so on and on. In this way, the backpropagation algorithm can modify the pattern of weights in very complex (multi-layer) artificial networks. In other words, it modifies the weights to each deeper layer of units so that at the same time it reduces the error on the output units (cf. Rumelhart, McClelland, and PDP Research Group 1986). A wide range of versions and extensions of the Hebbian and

2 Sigmoid functions are perhaps the most common, operating as a smoothed threshold function that is also differentiable. The fact that the activation function is differentiable is important in light of the fact that learning in an artificial network seeks to improve a performance metric that is evaluated via the activations state whereas learning itself can only operate on connection weights.

error-correcting algorithms have been introduced in the connectionist modelling literature (cf. Thomas and McClelland 2008).

The seventh feature of connectionist networks which has been emphasised particularly in the context of cognitive modelling is the representation of the elements and environment of the systems being modelled. Standardly, this is taken to consist of a set of externally provided events or a function for generating such events. For instance, a single pattern such as a visual input can count as an event, as well as a class of such patterns, e.g., the spelling of a word or its corresponding sound or meaning; or a sequence of inputs, such as words in a sequence. That is, as models of particular cognitive processes, ANNs are supplemented by a representational scheme which maps the elements of the cognitive domain of interest to the set of vectors depicting the relevant informational states or mappings for that domain. In what follows, I propose to explore which of these features of connectionist architectures bear on the individuation and explanations issue, respectively.

6.2.2 *The individuation of connectionist computations*

With these basic features of a connectionist architecture in hand, we can now consider the issue of the type-individuation of connectionist systems. As in the case of classical computationalism, I hold that this question can and should be treated independently of the question concerning the explanatory value of particular connectionist models of cognitive capacities. In brief, I take the individuation issue to be concerned with the characterisation of a particular type of structure (in this case connectionist networks) which may be applied to the study of various cognitive phenomena. In other words, the individuation issue focuses primarily on establishing the factors that play a role in the type-individuation of specific connectionist systems. The separability of the individuation and explanations issues allows a better grasp of the connectionist structures which are supposed to play an explanatory role in particular theoretical contexts by facilitating the investigation of the properties/features of connectionist networks which might play a crucial role in the modelling of a particular aspect of cognition.

In what follows, I analyse two general strategies of individuating connectionist systems and discuss how they differ from the formal or internalist individuation strategy of classical computational systems (cf. chapter 4). Whilst other individuation strategies might be proposed, I contend that these alternatives would raise the same or very similar issues to the ones discussed in connection with the two strategies sketched below. Hence, I will distinguish between two types of individuation strategies: (1) an architectural-based individua-

tion strategy, and (2) an input-output (wide) functional individuation strategy.

According to the architectural individuation strategy, the distinctions between the different types of connectionist systems (architectures) can be determined by considering three main features of the constituent elements of a connectionist system and their organisation: (i) their patterns of connectivity, (ii) the activation rules for processing units, and (iii) the learning algorithms used to train specific types of networks. As hinted above, the patterns of connectivity specify how the processing units of a network are connected to one another. According to this criterion, connectionist systems can be grouped in two major classes: (a) feedforward networks which have unidirectional connections, and (b) interactive networks which allow for bidirectional connections (cf. Bechtel and Abrahamsen 1991).

Feedforward connectionist networks standardly comprise a specific number of input units (or nodes), hidden, and output units organised into separate layers or levels, with units from one level feeding their activations forward to the units at the next level, until the final level is reached. The activation in such a network flows strictly from the input layer of nodes through the hidden layer to the output nodes (connections within a layer or from higher to lower layers are forbidden) in a way that depends on the specification of the initial values of the connection weights and the particular learning algorithm which is used to train the network. The simplest system with such a configuration will consist of only two layers of units: input and output units. When the connection strengths (weights) are properly set, even this type of very simple connectionist system can yield specific output patterns for different input patterns. However, since these two-layer connectionist networks have a limited computational power, feedforward networks have been developed to include more than two layers of processing units. Hence, the class of feedforward networks comprises both two-layer and multi-layered networks.

Further variations can be achieved in the class of two-layer and multi-layered feedforward architectures by letting, for instance, units at the same level send inhibitions and excitations to each other as well as to units at the next level up. An important variation of this type of architecture is the recurrent network (Elman 1990, 1991, 1993), which can receive input sequentially and modify its response appropriately depending upon what information was received at previous steps in the sequence. A simple recurrent network (SRN) can do this because, unlike a standard feedforward network, it is supplemented with a context layer that records a copy of the state of the hidden layer. This context layer feeds back into the hidden layer at the next time step. At any given point, the activation levels of the hidden units depend not only on the activation of the input units but also on the state of

these context units, which can be viewed as a sort of memory in this type of network.

Another class of connectionist systems that have been utilised in cognitive modelling are the interactive networks in which at least some connections are bidirectional, and the processing of any single input occurs dynamically across a large number of cycles (cf. Bechtel and Abrahamsen 1991). Unlike feedforward networks, interactive networks may or may not be organised into distinct levels. If they are thus organised, then it can be said that processing occurs both backwards and forwards. Examples of interactive networks include Hopfield nets (Hopfield 1982), Boltzmann machines, and harmony theory (Rumelhart, McClelland, and PDP Research Group 1988). Whilst none of these types of networks will be treated in detail anywhere in this chapter, for the purposes of discussing the individuation issue it suffices to point out that varying the organisation of connectionist networks alongside the dimension specified by the connectivity pattern yields distinct types of connectionist networks.

Even more fine-grained distinctions between connectionist systems can be achieved by taking into account not only their specific patterns of connectivity, but also the activation rules which determine the activation values of their component units after processing. In virtue of the type of activation values taken by individual units, connectionist systems can be distinguished into: (a) systems which take discrete activation values (which are typically binary), and (b) those which take continuous activation values which in turn can be unbounded or bounded. But the variations within the classes of connectionist systems can be further refined by taking into account the activation rules which specify how to calculate the level of activation for each unit at a given time. Although these rules can be quite similar for the two general classes of networks distinguished above (i.e., feedforward and interactive networks), they nevertheless yield a more fine-grained way of type-individuating different connectionist networks.

As outlined in the previous section, some of the most common activation rules are the linear activation rule and the logistic or sigmoid function (cf. Rumelhart, McClelland, and PDP Research Group 1986, 1988). Each of these activation rules can be adapted to obtain discrete rather than continuous activation values, typically for use in networks in which both input and output units are binary. In the case of the linear activation rule, the adaptation consists in comparing the net input to a threshold value. If net input exceeds the threshold, the output's unit activation is set to 1, otherwise being set to 0.[3] In the case

3 With a zero threshold, positive net input turns the output unit on and negative net input turns it off. A unit that uses a threshold in this way is called a linear threshold unit. For instance a network with an output layer of linear threshold units and an input layer of binary units is an elementary perceptron. Linear threshold units can also be used in the hidden and output layers of multi-layered feedforward networks and in interactive networks as well (cf. Bechtel and Abrahamsen 1991).

of the logistic activation rule, discrete activations can be achieved by using stochastic versions of the logistic function. When such versions of the logistic functions are used in a feedforward network of binary units, the relation between its net input and its activation becomes probabilistic. That is, the activation function determines the relative frequency with which the unit will turn on versus off.

The same activation rules that govern the propagation of activation in feedforward networks can be used in interactive networks as well. The only difference is that a special parameter t for time must be included as well because in these networks activations are updated many times on the same unit during processing. A further difference that affects the activation rules for interactive networks is that between synchronous update and asynchronous update procedures. In the first case, every unit's activation is updated once per timing cycle, whereas in the second case, there is no common sequence of cycles but rather a random determination of the times at which each unit separately is updated. Each update will require a separate application of the activation rule, unlike the case of the feedforward network where the activation rule is applied just once to each unit.

As this sketch already makes clear, there are a variety of ways of fixing both the types of activation a unit might take as well as the function which determines the unit's activation value. What all these functions have in common is that the new activation of a unit will depend to some extent on the net input the unit receives from other units in the network. The net input is determined in part by the weights on the connections which can be either hard-wired or can be determined by the network itself. This brings us to the last feature that plays a role in the architectural individuation of connectionist systems and that pertains to the distinctive ability of neural networks to change their own connection weights: learning principles.

As pointed out in section 2.1, learning for a connectionist system consists in changing the connection weights between simple units. Moreover, since the network is supposed to figure out the appropriate changes in weights without the intervention of an external programmer or internal executive in the network, the control over weight change must be entirely local. The learning principles should invoke or specify only information that can generally be available locally, such as the current activations of the units. This idea has inspired one of the first and most influential learning principles for connectionist networks, namely the Hebbian learning rule according to which whenever two units have the same sign, the connection between them is increased proportionally to the product of the two activations. Otherwise, the weight of the connection is decreased proportionally to the product of their activations (Hebb 1949). Numerous variations of the Hebbian learning rule can be used to train different connectionist systems. Further variation is introduced by training networks with

alternative learning functions such as the backpropagation algorithm and other related gradient descending methods which are said to increase the computational power of connectionist systems (cf. Elman 1996).

To sum up, from an architectural perspective, connectionist systems can be type-individuated in virtue of their patterns of connectivity, activation rules, and learning principles. This strategy will yield a large variety of connectionist systems, only some of which will be applicable in the investigation of particular cognitive phenomena. One advantage of this individuation strategy is that it captures many important features of the internal dynamics of a functioning connectionist network, thus allowing one to specify the relevant formal features that distinguish between different types of connectionist systems. This is important because some of these features (e.g., learning algorithms) have been taken to be essential for modelling specific aspects of cognitive phenomena, such as the gradualness of language learning and development or various interference effects in memory tasks.

Furthermore, the architectural individuation strategy comes very close to the formulation of the formal or internalist view of classical computational individuation defended in chapter 4. In the latter case, a system's computational identity is determined by the structural features that characterise its component elements, and their organisation, which in turn can be specified as a set of conditional instructions appropriately defined over the component elements of the system. Hence, the previous discussion supports the idea that, insofar as the individuation issue is concerned, there seems to be no significant difference between connectionist computations and symbolic computations.

One might nevertheless worry that this way of type-individuating various connectionist networks creates too great a gap between the formal features of connectionist networks and their application to the study of particular cognitive phenomena. Moreover, one may object that such a formal individuation strategy ignores the neurological inspiration underlying the original development of artificial neural networks. In order to address these worries, I sketch next another possible individuation strategy which seems to presuppose a more minimal gap between the individuation and explanation interests of connectionist modellers. I will return to the issue concerning the purported contributions of neurobiological constraints to the individuation of connectionist networks before concluding.

Another way to type-individuate connectionist systems is by the input-output functions they are able to perform or execute. In this case, the initial pattern of activation supplied to a connectionist network will be regarded as the relevant input that a system requires in order to yield, after a series of processing steps, a particular target

output. In this case, the connections are said to constitute constraints on the intended output and the stable state achieved by the system should be the state of the network that best satisfies these constraints. The adoption of this individuation strategy seems to be encouraged also by the fact that connectionists themselves insist on the centrality of the semantic interpretation of connectionist systems used in the study of different cognitive phenomena. That is, although they recognise that semantic interpretations are extrinsic to the connectionist systems themselves, connectionists also seem at times to imply that the type of input and output a given network is supposed to produce constrains the architecture of the system and thus its computational type. The idea implicit here seems to be that the semantic properties of the inputs and outputs processed by a given network determine (in part) its computational type.

Again, this latter interpretation of the connectionist functional individuation strategy is similar to some versions of the semantic view of computational individuation discussed in chapter 4 (e.g., Shagrir 2001). On this account, the semantic (externalist) contents of the inputs (and outputs) of a connectionist network are taken to *impact* its computational identity. This sort of strategy has been exploited by some defenders of radical connectionism (e.g., Frank, Haselager, and van Rooij 2009), but I will argue that it poses some serious problems for the applicability of connectionist models to the study of cognitive phenomena. However, as hinted above, there is also an internalist or formal characterisation of the functional individuation strategy which claims that connectionist systems are responsive only to certain formal features of the inputs/outputs on which the network operates (e.g., the patterns of activation corresponding to the input and output states, respectively). This is in turn consistent with the internalist view of classical computational individuation. Hence, whilst the wide (externalist) functional individuation strategy looks more appealing from a modelling perspective, I claim that it also tends to conflate the individuation and explanation issue, thereby compromising the independent characterisation of the connectionist models applied in the study of specific cognitive phenomena.

At a more general level, there are two considerations which reflect the main challenges confronting the input-output (functional) view of connectionist individuation. First, holding that the individuation of connectionist systems should proceed in terms of the input-output functions these systems are able to perform entails that connectionist networks are not that different from classical computational systems after all. For if two systems, one connectionist and the other classical, are able to compute the same input-output function, and furthermore they are individuated in terms of this particular criterion, then they will count as being computationally equivalent. The two systems might compute the same input-output function in different ways, but

from the point of view of their individuation, those differences will simply be irrelevant, because their type-identity will be determined by the input-output function they are able to perform (compute).

Second, the hypothesis that connectionist systems are individuated in functional terms is compatible with the idea that there exists a variety of networks which are able to perform the same input-output function. Otherwise put, the input-output functional individuation criterion seems to generate too coarse grained a taxonomy of connectionist systems. So, whilst the input-output functional individuation strategy might be viewed as a first step towards proving the applicability of connectionist networks to the study of cognitive functions or capacities, it also introduces an indeterminacy problem with respect to the connectionist system that would best capture the cognitive function being investigated. A possible way to further refine the individuation of connectionist systems postulated in cognitive modelling would be, as outlined above, to complement the functional criterion with a series of neurobiological constraints. These constraints would require that the connectionist systems postulated in the explanation of particular cognitive phenomena should be constrained by our current biological knowledge about the structure and organisation of the nervous system.

There are two straightforward problems with this proposal. Firstly, given the limits of our current biological knowledge concerning the structure and organisation of the nervous system and its components, it is reasonable to believe that the neurobiological constraints will be either too general or too indeterminate to fix in any strict sense the type-identity of any particular connectionist system. Secondly, the relevance of the neurobiological constraints for the individuation of specific connectionist systems rests on the strong assumption that all connectionist systems should model the particular patterns of connectivity of neurones in the brain or other neurobiologically relevant features such as the differences between various neurotransmitters that affect neural activity in different areas of the brain etc. Instead, I propose that connectionist systems which incorporate some neurobiologically based constraints constitute a sub-class of a wider variety of connectionist systems. This in turn implies that the type-individuation of connectionist systems cannot be determined solely in terms of neurobiological constraints either.

In addition, as it can be seen from any brief survey of the multiple connectionist models developed within the arena of cognitive modelling, the proposed architectures such as the LISA architecture (e.g., Hummel and Holyoak 1997, 2003), the neural blackboard architectures (van der Velde and de Kamps 2006), the tensor products based architectures (e.g., Smolensky 1990), and vector symbolic architectures (Plate 2003; Eliasmith and Stewart 2012) can be distinguished in terms of their architectural features along the lines sketched above.

This is mainly because neurobiological constraints are translated in the design of a connectionist system in terms of restraining the number of processing units, patterns of connectivity, activation rules, and learning algorithms that can be deployed by particular connectionist systems when embedded in a particular type of (cognitive) environment. Thus, whilst I would agree that neurobiological constraints play an important role in selecting the connectionist systems which are relevant for the study of particular cognitive phenomena, I maintain that they do not bear directly on the problem of connectionist individuation.

The preliminary conclusion of the preceding analysis is that an architectural individuation strategy is the most fit to capture the relevant features which play a role in the type-individuation of particular connectionist systems. In addition, there are good reasons to believe that such an architectural individuation strategy is very similar to the internalist view of classical computational individuation. The functional input-output individuation strategy, on the other hand, seems to be more convenient for identifying connectionist networks for specific modelling purposes. However, the discussion so far has left open the question of how to understand the practice of applying connectionist networks to the study of cognitive phenomena. In what follows, I seek to clarify this issue by discussing in some detail the representationalist problem as it arises in a specific connectionist setting.

6.2.3 *The representationalist problem*

This section tackles the issue of the semantic interpretation of connectionist systems. As will be shown below, this constitutes a crucial step in the applicability of connectionist networks to the study of cognitive phenomena. In section 2.1, I pointed out that an important feature of connectionist networks used in cognitive modelling is the representational scheme which provides a way of interpreting the elements of a connectionist network in a way that makes it relevant to the cognitive problem/phenomenon being modelled. Now I turn to analyse in more detail the role that these representational schemes play in the construction and development of good connectionist models of specific aspects of cognition.

At a more general level, the problem of the representational or non-representational character of connectionist approaches to cognition has been and continues to this day to be a major point of debate between defenders of classical computationalism, on the one hand, and proponents of connectionism, on the other hand (e.g., Fodor and Pylyshyn 1988; Smolensky 1988a,b; Jackendoff 2002; Bechtel and Abrahamsen 1991; Frank, Haselager, and van Rooij 2009; Eliasmith and Stewart 2012; Werning 2012, etc.). In what follows, I seek to clarify what is actually at stake in these sorts of debates and extract some

lessons that would allow us to advance our understanding of both classical and connectionist theories/models of cognition.

6.2.3.1 *A plea for internal structured representations*

Classicists claim that the explanation of certain salient properties of cognition (e.g., its productivity, systematicity, compositionality, inferential coherence, etc.) requires the postulation of internal structured representations, and rules appropriately defined over them. Connectionists maintain that cognitive processes and/or properties can be explained *exclusively* in terms of the transmission of activation patterns through the processing units (nodes) of highly complex neural networks. That is, according to connectionists, there are good reasons to believe that an explanatory theory of cognition should dispense with internal structured representations (symbols) and rules. Although it is not the case that all those who endorse connectionism would agree with this radical stance, I take it to be an assumption implicit in many presentations of the connectionist program in cognitive science (e.g., Christiansen and Chater 1994, 2008; Eliasmith and Stewart 2012; Werning 2012). The purpose of the following arguments is to show that a more moderate position is both available and more advisable.

Consider first the question whether connectionism ought to be viewed as a *representationalist* theory of cognition at all. The answer to this will depend of course on what you take a representationalist theory of cognition to be in the first place. Connectionists seem to be committed to characterise their theories as representational primarily in order to be able to qualify them as *proper* cognitive-level theories, as opposed to merely implementational ones. Connectionists accept that cognitive-level tasks/problems are standardly specified in representational or semantic terms. So there ought to be a way of connecting adequate connectionist theories of cognition with the representational/semantic descriptions of the phenomena being modelled. Along these lines, various authors have claimed that connectionist networks 'are explicitly concerned with the problem of *internal representation*' (cf. Rumelhart, McClelland, and PDP Research Group 1986, p. 121).

In response, a number of classical computationalists (e.g., Fodor and Pylyshyn 1988; Jackendoff 2002) have argued that this sort of representationalist commitment leads connectionists straight into the arms of a dilemma. For, they claim, representationalism goes hand in hand with the postulation of internal structured representations or symbols. In other words, according to this line of thought, the 'correct' way to understand the notion of mental representation or symbol is as something that must be able to participate in certain kinds of structured representations. But since most connectionists seem determined to reject a computational architecture that relies on symbols

and operations which are able to create novel internally structured symbols out of atomic symbols, Fodor and Pylyshyn (1988) conclude that connectionist theories/models should not be deemed to be *genuinely* representational in the first place.

A somewhat similar conclusion is defended by Ramsey (2007), who argues that connectionist theories are not representational in any 'interesting' sense. Ramsey agrees with other authors, such as Cummins (1983), Fodor and Pylyshyn (1988), and Jackendoff (2002), that assigning specific semantic values to the input and output nodes of a connectionist network does not suffice to qualify the system as being genuinely representational. He claims that, for a connectionist network to count as a proper representational system, the hidden layers and the learning algorithms must be interpretable in terms of structured representational relationships, on the model offered by classical computationalism. However, the latter requirement would conflict with the substantive connectionist claim that learning algorithms are general, domain-independent functions which operate on connection weights, changing the activation patterns between layers of simple processing units. Because of this, Ramsey (2007) concludes that connectionist models are essentially *non-representational*. Still, he points out that this verdict does not necessarily rule out the adequacy of connectionist models as tools for studying, and potentially explaining, various aspects of cognitive processing.[4]

I propose to read Ramsey's analysis as showing that connectionist models are representational only in a *weaker* sense. That is, admitting that semantic interpretations are actually external characterisations of computational models of cognitive processes, connectionist networks would count as weakly representational insofar as there exists a reasonable assignment of contents to their input and output nodes (cf. Ramsey 1999). There are two desirable consequences that recommend the cautious adoption of this proposal. First, this 'liberalisation' of the notion of mental representation is consistent with the idea that the adequacy of a computational model of cognition (classical or connectionist) does not depend exclusively on the availability of a direct semantic interpretation of *all* of its component structures and states. Second, a weaker (quasi-pragmatist) form of representationalism acknowledges the role that the selection of representational formats for the inputs and outputs of connectionist networks (models) plays in developing better models of cognitive phenomena. Therefore, a weakening of the representational thesis would allow us to focus more on the constraints that determine the choice of particular representational schemes used by different connectionist modellers and to understand how they contribute to the potential explanatory

4 Although it is not clear how Ramsey would account for the adequacy of connectionist theories of cognition, his solution to the classicist's dilemma may be viewed as a reasonable proposal which at least avoids spurious debates concerning the meaning of the term 'representation'.

power of the proposed connectionist models. Before considering this issue in more detail, I would like to say something about where this line of argument leaves the objection raised by classicists against connectionism's inability to explain certain salient features of cognitive processes such as their productivity, systematicity, compositionality, and inferential coherence.

6.2.3.2 *Connectionist alternatives*

Consider first some of the major response strategies that have been put forward in the connectionist literature. One of the most radical response routes defended by a number of connectionists is the so-called *approximationist approach* (cf. Rumelhart, McClelland, and PDP Research Group 1986; Smolensky 1988a; Bechtel and Abrahamsen 1991). On this view, connectionist theories will one day provide the most detailed and accurate account of cognitive processing. Classical symbolic models, on the other hand, are viewed as more abstract (idealised) accounts which nevertheless manage to capture in an adequate way certain important features or patterns in the cognitive phenomena being investigated. It is in this sense that symbolic models are said to *approximate* connectionist models, viz. by providing a less detailed account of the target cognitive process/phenomenon than the connectionist model.

Although the term 'approximation' is somewhat misleading, the driving thought of this connectionist strategy seems to be the following. Classical rule-based computationalist theories of cognition cannot capture adequately 'the variability, flexibility, and subtlety' displayed by cognitive phenomena, because the rules used to account for such phenomena are 'brittle'. Hence, connectionists argue, structured symbols and representations might be useful in constructing higher-level models that abstract away from many of the details extracted from the behavioural data, but in order to model the *actual* mechanisms underlying various cognitive processes, more detailed and less brittle models are needed. This is taken to imply that an explanatory theory of cognition will not require the postulation of internal structured representations and rules and instead it will specify, in a broadly connectionist framework, the mechanisms underlying various cognitive capacities and/or processes.

The standard way of arguing for the approximationist approach consists in pointing out that there is a wide class of connectionist models or simulations currently being developed which can accomplish cognitive tasks that would seem at first blush to require the use of internal symbols and rules. In fact, this 'exemplification' strategy is used on many occasions to support the appeal of connectionist approaches to cognition. The general argument is that such connectionist simulations/models provide something like an inductive basis by showing how an 'approximation' of a cognitive task might be car-

ried out by a very simple connectionist system. Then, assuming that the behaviour displayed by the connectionist network will scale up to accommodate realistic demands on complex cognitive systems, one should be inclined to conclude that a purely connectionist account of the mechanisms underlying cognition is in principle possible precisely in virtue of this sort of 'inductive proof'. Whilst some of these models will be analysed in more detail in section 3 below, I would like to make here two quick remarks about the implications of this connectionist strategy with respect to the 'classicist's dilemma'.

First, it should be noted that in order for this sort of inductive proof to work at all, one must be able to guarantee that the connectionist models are not themselves relying on implicit rules that determine their 'successful' performance. In fact, there are a number of studies which have shown that at least some of these successful connectionist models *do* rely on the implementation of hidden classical rules (e.g., Marcus 2001; Ramsey 2007). However, these criticisms should not be taken as providing an impossibility proof for the construction of non-rule-governed connectionist systems that 'approximate' certain well-specified cognitive tasks. Actually, they can be interpreted as highlighting additional constraints that connectionist models must satisfy if they are to count as proper non-classical accounts (cf. Eliasmith and Stewart 2012).

Second, most connectionists who endorse the approximationist approach take it to entail that the properties singled out by classicists (e.g, the productivity, systematicity, etc. of cognition) are not actually instantiated by real cognitive systems. So they take successful connectionist models to illustrate how more limited versions of these properties might be exhibited by actual cognitive systems but deny their presence in cognitive systems generally. Otherwise put, the approximationist approach partly rejects the classicist challenge to explain in a purely connectionist way the productivity, systematicity, and other similar properties of cognition because these properties are viewed as idealised (abstract) versions of the properties displayed by actual cognitive systems (e.g., Christiansen and Chater 2002; Christiansen and Chater 2008; Werning 2012).

Another type of response to the classicist challenge that has been discussed in the connectionist literature is the so-called *compatibilist approach* (Touretzky and Hinton 1988; Dyer 1991). According to this view, there are certain cognitive processes such as complex reasoning, decision making, and problem solving that could not be explained without positing internal symbols that are systematically manipulated by rules. This account is then consistent with the idea that there might be certain features such as productivity, systematicity, and coherence which characterise precisely those types of cognitive processes that are inexplicable without the appeal to internal structured representations and rules. But connectionists should not even

attempt to take up the challenge to explain these patterns/features in the first place because they are appropriately stated in terms of the relations that hold between the terms posited by the symbolic description of the cognitive processes. Under this view, the purpose of connectionist models is to provide plausible *implementational* accounts for the rule-governed cognitive processes.

To further clarify the previous claim, it should be pointed out that, unlike the approximationist approach, which proceeds in a bottom-up way by showing that some appropriately trained networks can exhibit certain types of regularities that correspond to patterns observed in cognitive processing, the compatibilist approach works in a top-down fashion. That is, proponents of the compatibilist approach (e.g., Touretzky and Hinton 1988) usually start with the rule processing description of a given cognitive tasks and then try to construct a network that would be able to implement the rules in question. However, despite the acknowledgment that some connectionist systems provide an implementational rather than cognitive-level account of certain cognitive phenomena, the compatibilist approach should not be taken to imply that an implementational characterisation diminishes the importance of having a connectionist account of a particular cognitive phenomenon.

Thus, *contra* classicists like Fodor and Pylyshyn (1988), compatibilists do not regard connectionist models as being in any way 'inferior' on the grounds that they are 'mere' implementational accounts of a given cognitive capacity. Although the arguments developed by these authors are not fully articulated, the implicit claim seems to be that such implementational connectionist models can reflect important computational properties that would not be tractable in classical symbolic implementations. This argument is a potentially important challenge to Fodor and Pylyshyn's (1988) contemptuous characterisation of implementational accounts of cognition, for it would show that connectionist models may even contribute to the development and refinement of better cognitive-level classical symbolic models.

Finally, another way to account for the relation between connectionist architectures and the notion of internal structured representations and rules consists in saying that networks may develop the capacity to interpret and produce symbols that are external to the network itself. This alternative has been put forward by Rumelhart, Smolensky, McClelland, and Hinton (1986) and subsequently elaborated by Smolensky (1988), Clark (1997), and more recently, illustrated by the connectionist model of sentence comprehension constructed by Frank *et al.* (2009). The key idea of this approach is that the internal structures postulated by a classical computationalist approach are stable patterns that a network can acquire in time by sufficient exposure to a highly structured environment. In other words, the internal symbol-and-rules structures of certain cognitive capacities are said

to be the result of a complex interaction between certain connectionist networks and a 'symbolically structured' external environment (Frank, Haselager, and van Rooij 2009).

According to this strong 'externalist' approach, the properties highlighted by classical computationalists (productivity, systematicity, etc.) are extracted by connectionist networks from a highly symbolically structured external environment. This way of explaining these particular features of cognition has the purported advantage of allowing for imperfections in the ways in which actual cognitive systems under different conditions display these particular properties. That is, the productivity, systematicity, and coherence of cognitive systems is *prima facie* limited by the degree to which these properties are present/instantiated by the external environment in which a particular cognitive system is embedded. The most challenging part of this approach is to account for the way in which external symbolic structures might be internalised (to some extent at least). Assuming for the sake of the argument that such an account would be readily available, it is important to note that, on this approach, the internalised symbols are simply patterns in a network. That is, the network is said to settle, via some form of dynamic encoding (which is arguably quite distinct from the construal of symbols in the classical framework) into certain stable states which count as 'symbols'. Thus, if one brackets the reservations expressed above, the externalist approach would seem to afford a purely connectionist explanation of the salient features highlighted by classicists.

I take the preceding discussion to have shown that, contrary to the strong classicist contention, connectionists have available a number of strategies for avoiding the challenge that neural networks are not an appropriate tool for investigating interesting cognitive phenomena. Whilst the availability of all these strategies tends to suggest that the dilemma posed by classical computationalists is less definitive than standardly assumed, none of the connectionist responses is free from criticism. As outlined above, the approximationist strategy implies that the properties emphasised by symbolic approaches are in fact mere theoretical artefacts and that the actual features which are in need of explanation are much more fuzzy and vague. Whilst connectionists welcome this consequence and claim that it is a way of reconceptualising certain properties of cognitive phenomena, it can be shown that the strategy endangers the notion of connectionist explanation itself. That is, the bottom-up strategy promoted by defenders of an approximationist strategy makes it look like the explanandum is defined as whatever can be generated by an appropriately trained connectionist network, rather than something which is a robust feature of the phenomenon/system being modelled with the help of connectionist system. Not only does this seem to reverse the order of explanation, but it also puts into question what counts as an

appropriate training regime for the connectionist model in the first place.

The 'externalist' symbolic approach also faces a series of challenges, such as that of specifying the mechanisms via which connectionist systems can settle into stable internal structured states and the way in which these are then used at different stages of cognitive processing. Without such an account, the plausibility of the approach seems to rest again on the success of particular networks in extracting some structure from a carefully manipulated/constrained environment. Although a number of such examples have been discussed in the literature, the inductive proof which is supposed to secure the generality of the approach is vulnerable to the fact that the success of these models depends to a large extent on the control of the external environment from which the network is meant to extract a particular type of symbolic structure. Realistic environments contain much more noise, so an appropriately scaled-up network is very likely to perform sub-optimally in such an unconstrained setting. The compatibilist approach seems to avoid the classicist challenge, but it is unclear where it actually stands with respect to the idea of genuine classical symbolic explanations of cognitive phenomena.

Because of all these limitations, I propose to treat the connectionist strategies as non-exclusive alternative ways of responding to the classicist dilemma. This suggestion implies that connectionist models/simulations are apt to play not one but a host of different roles in cognitive modelling. Moreover, this proposal allows that, for certain aspects of cognitive processing, connectionism might provide a more adequate treatment than a classical rules and representations-based account. For instance, connectionism seems well positioned to explain certain 'unexpected' or rare cognitive effects which arise in various domains of cognitive processing, such as the catastrophic effects in learning and memorisation tasks, or certain types of illusions in visual processing and perhaps also various types of developmental disorders (e.g., McClelland et al. 2010).

However, this aspect of connectionism does not support the ambitious hypothesis that connectionism will one day provide a unified theory of cognition. In particular, I contest the connectionist claim that all symbolic models should be treated as heuristics for better and more detailed connectionist theories of cognition. As in the case of 'mere' implementational treatments of connectionism, I think the heuristic qualification is misleading in creating an artificial polarisation between classical and connectionist approaches to cognition. Instead, I maintain that both frameworks make available important ways of conceptualising and explaining different aspects of cognitive phenomena. In what follows, I briefly return to discuss the role of representational schemes in connectionist explanations of cognitive phenomena. This discussion complements and extends the notion of a

'weak' representationalist character of connectionist models/theories of cognition.

6.2.4 *Representational schemes and connectionist explanations*

The structure of this section comprises two interrelated parts. I start by briefly analysing the distinction between *localist* and *distributed* representations, and then turn to the question of how semantic interpretations of connectionist systems contribute to the evaluation of the explanatory power of connectionist models of specific cognitive phenomena. In a nutshell, on the *localist* approach to the semantic interpretation of connectionist systems, each concept from the relevant domain of the cognitive problem being modelled is assigned to one distinct unit of the network, whereas on the *distributed* approach, each of the relevant concepts are distributed across multiple units. Each of these approaches has its own advantages and limitations.

The primary advantage of localist models is that units can be labeled in the most intuitive way possible so as to facilitate keeping track of what the network is doing. The disadvantage of this strategy is that one might be tempted to assume that the semantic interpretations assigned to individual nodes of the network are intrinsic, i.e., they convey further (semantic) information that contributes to the processing of the network. However, as was shown in the discussion of the issue of connectionist individuation, there is no *prima facie* reason to believe that semantic contents contribute to the individuation of different connectionist systems. Hence, the correct way to conceive of the process of semantic interpretation is that it constitutes an external (pragmatic) step which establishes a correspondence between the concepts that define a particular cognitive problem and the elements of the connectionist system that is supposed to model that particular cognitive problem. This further implies that when a particular network is being used as a model of a particular cognitive task, its success will depend to a significant extent on the modeller's intuitions about the ways in which the relevant concepts in the target cognitive domain are interconnected and how they should be encoded (i.e., associated with the units of the network), as well as on her skill in setting up connections appropriate to that encoding.

The 'extrinsic' task of semantic interpretation is even more complex under the distributed approach. For in a distributed network, each concept is represented by a pattern of activation across a set of units. That is, on such an approach no single unit can be said to represent a concept on its own (i.e., individual units do not have semantic interpretations). One way to achieve such a distributed representation consists in carrying out a so-called *featural analysis* of the concepts that characterise a given cognitive task, and then encode that analysis across an appropriate number of units. As such, feat-

ural representation actually involves a localist representation of the extracted features (one feature per unit), and a distributed representation of each of the relevant concepts. There are two main ways to obtain a featural analysis of a target cognitive domain which would generate the distributed representations used by a network to carry out a particular cognitive task.

One option is to rely on an existing theory to guide the featural analysis; for instance, a particular linguistic theory may be used as a basis for representing words as patterns over phonemic units. Although this method suggests a close connection between connectionist and classical accounts of cognitive phenomena, connectionists insist that since the semantic interpretation is extrinsic to the functioning of the connectionist network *per se*, the essential contribution made by networks in a cognitive modelling context is to show how a connectionist cognitive system might make use of such a distributed representation in order to carry out a particular task. That is, the explanatory contribution of the network would consists in showing something about the mechanisms underlying the cognitive phenomenon/task being modelled. Still, there is a sense in which the (higher-level) theory used to derive the featural analysis provides a series of constraints on the mechanisms and patterns explored with the help of the connectionist system. That is, the theoretical description can be said to secure the possibility of applying a connectionist system in order to study certain aspects/patterns of a given cognitive phenomenon. There is nothing in this *compatibilist* picture that precludes the possibility of the connectionist system being used to reveal features or patterns in the target cognitive phenomena which are not tractable from an alternative (e.g., classical computationalist) perspective.

Another option is to let the system perform its own analysis of the domain. In this case, when a network is trained with a particular learning algorithm, the modeller specifies only the semantic interpretations for the input and output units; in addition, the input-output cases used for training are selected with respect to this particular interpretation. However, under this method of carrying out the featural analysis, the modeller can remain agnostic about the aspects of the input-output cases which each hidden unit will become sensitive to. The learning process can be viewed as a process of feature extraction which assigns to the hidden units so-called 'intermediary representations' which can nevertheless fail to correspond to the most salient regularities (features) in the input. Instead, each hidden unit is said to be sensitive to complex and subtle regularities called *microfeatures*. On this account, each layer of hidden units can be viewed as providing a different distributed encoding of the input pattern.

Connectionist modellers are usually keen to emphasise a number of advantages of this method of obtaining the featural analysis which de-

termines the distributed representation of the concepts that define a particular cognitive task. Namely, they claim that, under this method, partiality of the representation is still compatible with successful performance. The general point is that once information is distributed as a pattern across units, the resources used to process it are also distributed, which in turn implies that a system using this information would be more resilient to damage. In addition, this method is said to allow the system to learn new information without sacrificing existing information. For instance, without adding new units, one can teach a network to respond to new input that is significantly different from any it has encountered until a certain point simply by making slight changes to a variety of weights so that one does not affect the way in which the network responds to existing patterns.[5] In addition, connectionists emphasise that distributed representations are a good way of imposing a series of quantitative constraints on connectionist models so that they may be viewed as neurologically plausible in that they would not require huge processing resources that a human brain cannot support.

The preceding considerations suggest that, as in the case of classical computationalist models, semantic interpretations, whilst extrinsic to the connectionist systems themselves, play an important role in evaluating the explanatory power of connectionist models of particular cognitive phenomena. That is, they connect certain connectionist systems to particular cognitive problems or tasks that the system is meant to perform. The performance of the system in turn is taken to provide some insight into the functioning of the cognitive system being investigated. However, unlike the case of classical models where semantic interpretations are facilitated by the classical decompositional strategy which maps simpler computational elements into simpler semantic interpretations and more complex computational elements into appropriately structured semantic interpretations, both localist and distributed representational schemes fail to reflect this type of symmetry. As we saw above, the semantic interpretation of hidden units and of learning algorithms is at best partial and approximate. This is in part what makes purported connectionist explanations of particular aspects of cognitive processes harder to grasp.

However, the fact that connectionist explanations seem to be epistemically more complex need not be taken as a sign that they are not *proper* cognitive-level accounts. I think that connectionist explanations are indeed better viewed as providing an alternative to both classical computational and mechanistic explanatory strategies. This is primarily because network architectures attempt to explain particular aspects of cognitive phenomena without assuming the sort

5 It should be pointed out that there are limits to this capacity, i.e., in certain circumstances when the system has to learn new inputs, it can display so called 'catastrophic interferences' which mark the fact that the new input disrupts previous learning to an unacceptable degree.

of componential decomposition that characterises both the classical symbolic and the mechanistic approach (cf. Bechtel and Richardson 1993/2010). Instead, connectionism conceives various cognitive effects and/or processes as the result of the complex organisation and interaction of a system of simple units governed by simple updating (i.e., learning) rules. Precisely because most connectionist models/theories bypass decompositional and localisation concerns, they are not strictly speaking implementational accounts but rather abstract (i.e., mathematical) ways of describing and elucidating certain aspects of cognitive processing/performance. As such, they are *in principle* consistent with other strategies of investigating and explaining cognitive phenomena.

In summary, when considering the issue of the explanatory value of connectionist models/theories of cognition, I distinguish between two questions: (i) what makes connectionist systems potentially explanatory of cognitive phenomena, and (ii) what sort of factors contribute to the construction of better explanatory connectionist models of specific cognitive phenomena. The answer to the first question, sketched above, is that connectionist systems, structurally or formally individuated, have a series of features which reflect certain important patterns or features of cognitive phenomena. Otherwise put, they allow us to conceptualise certain cognitive phenomena in ways that advance our understanding of them. Furthermore, when used in conjunction with other explanatory frameworks, they can yield novel hypotheses about the structure and organisation of cognition at different levels of abstraction, e.g., symbolic and/or neurobiological. As for the second question, I think it is better approached from a practice-based perspective which would afford a more fine-grained analysis of the factors that contribute, in the current cognitive modelling practice, to the development of better connectionist explanations of cognition.

6.3 CONNECTIONISM FROM A PRACTICE-BASED PERSPECTIVE

Nowadays, connectionist models are applied within most of the subbranches of cognitive science, and, in certain cases, have been taken to yield explanatory hypotheses about certain surprising aspects of cognitive processing (e.g., the catastrophic interference effect in memory tasks, cf. McClelland et al. 2010). In what follows, I will focus on two models proposed in the domain of language processing: (i) a model of English past tense inflection (Rumelhart, McClelland, and PDP Research Group 1986), and (ii) a model of word recognition (Elman 1996; Elman 1990).[6] The aim of this critical analysis is twofold. Firstly, I seek to establish which architectural features of these partic-

6 This choice of examples is primarily motivated by the fact that the great majority of connectionist models which have been singled out for the purposes of philosophical analysis belong to this area of cognitive research.

ular models are and are not essential to their functioning. In other words, I investigate the specific architectural commitments that are taken to make connectionist models adequate descriptive and potentially explanatory tools in the investigation of cognitive phenomena. Second, I show that these commitments are consistent with the fact that certain classical principles and/or hypotheses are implicitly used in at least some 'successful' cases of connectionist modelling. More generally, I maintain that PDP modelling is compatible with a range of classical hypotheses about the nature and structure of cognition. In the last section of the chapter, I discuss in more detail the consequences of this proposal. For now, I turn to the connectionists models themselves in order to clarify how they work and what they show about the general structure of connectionist explanations of cognitive phenomena.

6.3.1 *Connectionist models of linguistic inflection*

Among the various cognitive problems to which connectionist modelling has been systematically applied, linguistic inflection is perhaps the one that has generated most interest and controversy in philosophical circles. More specifically, the acquisition of English past tense has often been taken to illustrate the substantive clash between classicism (i.e., generative linguistics) and connectionism (cf. Marcus 2001; Yang 2002). Although assigning a central role to the problem of past tense in the study of language can be misleading, for present purposes it suffices to assume that it constitutes an instructive example for assessing the relationship between the two theoretical paradigms.[7] The analysis pursued in this section comprises two steps. First, I sketch a minimal background for understanding the empirical and theoretical context in which connectionist models of English past tense acquisition have been developed. Second, I identify the main architectural commitments of this class of connectionist networks and discuss their explanatory function.

The problem of past tense concerns the systematic patterns observed in children's acquisition of past tense. For instance, in English children (and adults) in general inflect novel verbs with the *-d* suffix, as in *cross-crossed*. Another important fact is that young children sometimes *overregularize*. That is, they sometimes produce patterns such as *take-taked* instead of *take-took*, where the suffix *-d* for regular verbs is used for an irregular verb. According to the most extensive study of past tense acquisition (Marcus et al. 1992), overregularization occurs on average in 10% of all instances of irregular

7 A number of authors have pointed out that 'the problem of past tense, particularly in English, notorious for its impoverished phonology, is a marginal problem in linguistics, and placing it at the centre of attention does no justice to the intricacy of the study of language' (cf. Yang 2002, p. 59).

verbs. In addition, errors such as *bring-brang* and *wipe-wope*, that is, mis-irregularization errors where children misapply and overapply irregular past tense forms, also occur but are very rare, representing about 0.2% of all instances of irregular verb uses (cf. Pinker 1995).

The leading connectionist approach to the problem of English past tense inflection claims that the systematic patterns noticed in past tense acquisition emerge from the statistical properties of the input data presented to connectionist networks (cf. Rumelhart, McClelland, and PDP Research Group 1986). To unpack this claim, consider first the influential Rumelhart and McClelland (1986) model of past tense inflection. The original model works by taking a phonetically encoded input and transforming it into a phonetically encoded output. For instance, the input to the model on a given trial is a phonetic description of the word *sing*, while the target output is *sang*. On this model, words consist of sets of triples, known as *Wickelfeatures*. Thus, a word like *sing* is represented by the simultaneous activation of the triples *#si, sin, ing*, and *ng#* (where # is a special marker for the beginning or end of a word).

Although the model initially proposed by Rumelhart and McClelland (1986) did not have any hidden units, it performed surprisingly well at capturing some interesting qualitative phenomena. For instance, even if the model did not implement any explicit *-d* rule, it was able to extend the default inflection to a number of novel verbs (i.e., not present in the training space), yielding overregularizations such as *breaked* or *taked*. Also, the model was able to correctly inflect a number of irregular verbs before it first began to overregularize.

Despite the initial enthusiasm triggered by the performance of this model, it is now widely acknowledged that the model is seriously flawed. For example, it has been shown that the feature mentioned above, i.e., the ability of the model to capture a period of correct irregular use prior to overregularization, was dependent upon an unrealistic, abrupt change from an almost irregular input vocabulary to an almost entirely regular input vocabulary (cf. Marcus 2001). Another more serious problem is that the model yields bizarre blends such as the past tense *membled* for *mail* and the past tense *imin* for the novel verb stem *smeeb* (not encountered in acquisition data). In addition, it turns out that the Wickelfeatures system used to represent words on this model cannot keep certain pairs of words distinct which in turn affects the statistical distribution of the input from which the network is supposed to learn reliably past tense inflection (cf. Pinker and Prince 1988; Marcus 2001).

Nevertheless, the initial wave of criticism against Rumelhart and McClelland's model of linguistic inflection left open the question why the model had failed to appropriately accommodate the empirical data in the first place. Whilst a number of authors (e.g., Pinker 1995; Marcus 1999) have attributed the model's limitations to its lack of

rules, others have argued that their source lies in the fact that the model lacks a hidden layer. Other connectionists have proposed that the use of the backpropagation algorithm (or other gradient descent learning functions) in a network with hidden units constitutes an important step forward in the application of ANNs to the study of cognitive phenomena such as linguistic inflection (cf. Hare, Elman, and Daugherty 1995). In consequence, various researchers have proposed more sophisticated networks, which are enhanced with a hidden layer and more plausible training regimes, as well as phonetic representation schemes.

However, it should be noted that the continuity between these newer models and the original model of past tense inflection is non-negligible, for they all treat the task of past tense acquisition as one of using a single network for learning 'a mapping between a phonologically represented stem and phonologically represented inflected form' (cf. Marcus 2001, p. 72). Moreover, the general aim of all these inter-related models is to show that 'regular and irregular verbs ... [are] represented and processed similarly in the same device' (cf. Elman 1996, p. 139).

A first pass at a general criticism of the connectionist approach to the phenomena of linguistic inflection pertains to the bold 'unificationist' aim invoked by Elman (ibid.). For if connectionist principles were sufficient for the explanation of the patterns observed in the acquisition of English past tense, it is rather surprising that there is as yet (more than 30 years later) no substantial proposal of a comprehensive single-mechanism model for English past tense inflection. Instead, the models currently available are targeted at different aspects of the past tense formation: e.g., one model for why denominal verbs (such as *ring* as in *ring a city with soldiers*) receive regular inflection (Daugherty et al. 1993), another for handling defaults for low frequency verbs (Hare, Elman, and Daugherty 1995), another for distinguishing homonyms that have different past-tense forms (MacWhinney and Leinbach 1991), and still another for handling overregularization phenomena (Plunkett and Marchman 1993a,b). This fragmentation of the connectionist modelling space seems even more problematic when one takes into consideration the fact that existent models differ from one another in their representational formats for the input and output units of particular networks, as well as in their training regimes. As such, the prospects of a single integrated connectionist model for learning past tense seem to be quite bleak. Contra Elman's (1996) optimistic pronouncement, rather than showing how inflection can be achieved by a single device, current models seem to suggest that more than one mechanism is necessary for linguistic inflection.

Another general difficulty facing these models is that they do not match up very well to the quantitative data gathered from both child language acquisition and comparative studies of the world's languages

(cf. Marcus 2001; Yang 2002). This is a serious challenge especially given the danger highlighted above. Connectionist accounts tend to reverse the order of the explanation in the sense that they are more concerned to justify and motivate empirically the patterns generated by the performance of certain connectionist networks rather than to show how certain salient (problematic) patterns in the cognitive phenomena are accounted for by the principles governing the functioning of such systems. Moreover, this problem persists even if one grants that connectionist modelling is still in its infancy and this somehow justifies the quantitative limitations of existing connectionist models/simulations of cognitive functions.[8]

Although current connectionist models of linguistic inflection do not provide a unified account of the cognitive task they were intended to model, and although they fail to meet other adequacy constraints such as developmental compatibility and quantitative constraints from language acquisition and cross-linguistic studies, they can nevertheless be taken to provide appropriate (statistical) tools for modelling the the gradualness of language learning in the particular domain of English past tense acquisition. Otherwise put, whilst these models are not decisive in ruling out a classical rule-based approach to language acquisition, they have arguably facilitated the investigation of a number of learning principles (paradigms) which characterise the gradualness of the cognitive process of linguistic inflection.

That is, these models suggest that some of the mechanisms that are responsible for the acquisition of English past tense involve the sort of general principles reflected in the functioning of the connectionist systems used to model this particular cognitive task. Whilst limited, this contribution elucidates a feature of cognitive processing which was not directly tractable from a classical rule-based perspective. Thus, even if we were to discover that future, perhaps hybrid, models of English past tense inflection are able to capture the phenomena in more detail and satisfy to a greater extent both quantitative and other types of adequacy constraints, I claim that this would not cancel out completely the distinctive explanatory role played by connectionists models in the investigation of this complex cognitive phenomenon.

8 For instance, a number of connectionists have argued that the objections raised against the original model proposed by Rumelhart and McClelland (1986) can be avoided by using an alternative class of networks called classifiers (e.g., Hare, Elman, and Daugherty 1995). For a detailed criticism of the latter class of models as providing a radical alternative to classical symbolic accounts of linguistic inflection, see Marcus (2001). The basic idea of Marcus's argument is that the successful performance of classifier models is due to the fact that they rely on two implicit classical rules/operations performed by the external clean-up network included in Hare *et al.*'s (1995) connectionist architecture.

6.3.2 *Connectionist models of word recognition*

In what follows, I propose to investigate the architectural commitments of yet another class of connectionist networks that are offered as an alternative to classical models that postulate rules and complex structured representations. It is worth pointing out from the outset that although this class of models was originally developed for the task of word recognition, it has been very influential in other areas as well, such as acquisition of syntax and phonology, where it is viewed as challenging more directly the classical generative framework. Broadly, these models are taken to demonstrate how the linear organisation of neural networks can give rise to apparently intricate structural dependencies.

Word recognition is an instance of the more general problem of extracting structure from what appears to be linear input data. The most influential approach to modelling this task in connectionist terms has been developed by Elman (1990, 1991, 1993). The architecture proposed by Elman (1990) is known today as the simple recurrent network (SRN). As pointed out in section 2, a SRN has, in addition to the input, output, and hidden layers, a layer of context units to which a copy of the activation of the hidden units at given time step t is transmitted. Subsequently, this activation vector is fed back to the hidden units at time $t + 1$. Standardly, such a network is trained with a version of the backpropagation algorithm, just as any other multilayer feedforward network, ignoring the origin of the information coming from the context layer. The latter is simply meant to guarantee that each input to the SRN is processed in the context of what came before. This feature of the network is supposed to permit the network to learn statistical relationships across sequences of inputs, i.e., to extract structure from a wide range of available inputs. Part of the appeal of this type of connectionist model comes from its robust applicability to a diversified class of language processing problems.

Across a series of applications, the use of this type of connectionist networks was taken to show that the internal representations extracted by connectionist networks from input data are not necessarily 'flat' but could include hierarchical encodings of category structure (cf. Elman 1996). A connectionist model that has been taken to illustrate this hypothesis is the one discussed by Elman (1991). Elman focused on 'long-range dependencies', which are links between words that depend only on their syntactic relationships in the sentence (and not on their separation into words, for example). One instance of this sort of relation is subject-verb number agreement (as in *the boy chases the cat*, but also in *the boy whom the boys chase chases the cat*).

Using a SRN trained on a prediction task similar to the original one (ibid.), in which the network was supposed to predict the next letter in a series of concatenated words, Elman (1993) showed that the net-

work was able to learn certain types of long-range dependencies, even across the separation of multiple phrases. For instance, if the subject of the sentence was *boy*, when the network came to predict the main verb *chase* as the next word it also predicted that it should be in the singular. It is worth considering briefly the way in which the performance of the network was subsequently evaluated. Elman explored the similarity structure in the hidden unit representations, using principal component analyses to identify the salient dimensions of similarity across which activation states were varying. This allowed the reduction of the high dimensionality of the internal states to a manageable number that afforded a better visualisation of the underlying process. In this way, he was able to plot the trajectories of activation as the network altered its internal states in response to each subsequent input. Following the trajectories across different dimensions of the principal component space, he was able to show that the network adopted similar states in response to particular lexical items but also that it modified the pattern slightly according to the grammatical status of a word. This was then taken to suggest that some of the emerging trajectories might encode the singularity/plurality feature, whereas others might be responsible for different types of structural features.

The representation of sentences by trajectories through an activation space in which the activation pattern for each word is subtly shifted according to the context of its usage was taken to challenge the classical view according to which symbolic structures (possessing fixed type-identities) are bound into particular grammatical roles by specific syntactic constructions. Moreover, in light of the same preliminary results, Elman and his followers proposed that the property of compositionality at the heart of the classical symbolic computational approach is not required to explain language processing after all (cf. Christiansen and Chater 2002, 2008). Instead, a new series of categories, such as context-sensitive compositionality and restricted (*ersatz*) recursivity, are better suited to characterise and explain the salient patterns observed in language production and comprehension.

Two limitations of the SRNs used to model various aspects of syntactic acquisition are particularly salient. Firstly, in the proposed simulations, the prediction task does not seem to learn any sort of categorisation over the input set (cf. Marcus 2001). Although the simulations demonstrate that information important for language production and/or comprehension can be induced/extracted from word sequences, neither task is actually performed. This seems to be the case since the learned distinction between the categories of noun and verb emerging in the hidden unit representations is completely tied up with carrying out the prediction task itself. However, in order to perform a task such as language comprehension, the SRN would need to

learn categorisations from the word sequences. That is, the network must be able to decide, for instance, which noun was the agent and which noun was the patient in a sentence, regardless of whether the sentence was presented in the active or passive voice. These types of computations are more complex and the available connectionist solutions are typically problematic from a tractability perspective.

Secondly and more generally, whilst connectionism tends to be promoted as a radical and unificatory account of linguistic phenomena, the existent connectionist models are very far from providing an integrated picture of language acquisition and/or development. As also suggested by the previous example, I think that this is in part because connectionism has not really tackled the problem of language acquisition in a broad enough empirical context. As illustrated with the help of the models analysed in this section, a main line of the connectionist research is dedicated to proving that specific neural networks are able to capture some limited aspects of language processing. Whilst these results might be revealing in certain respects, I think that they do not suffice to support the strong claim that connectionism constitutes the single most appropriate framework for studying cognitive phenomena. That is, the preceding analyses shows that the appropriately circumscribed explanatory value of particular connectionist models of cognition is compatible with there being other explanatory frameworks (except connectionism) that are appropriate for the investigation of cognitive phenomena.

6.3.3 *The allure of the connectionist approach*

Given the challenges facing current connectionist models, and the difficulty of justifying their specific/distinctive explanatory contributions to the understanding of cognitive phenomena, it is no surprise that connectionists have invoked a series of more general considerations in support of the adoption of neural networks as an adequate tool for the investigation and explanation of cognition.[9] In what follows, I discuss briefly two of these more general considerations in order to shed more light on the conclusions concerning the nature and structure of connectionist explanations of cognitive phenomena.

9 This move has attracted even more criticism from defenders of classical computationalism who have pointed out that the master argument against classical computationalism does not seem to rest so much on empirical arguments against particular classical models of specific cognitive capacities, but instead on theoretical considerations such as those sketched in this section concerning general features of connectionist networks (e.g., Marcus 2001; Gallistel and King 2009). Whilst I agree with the classicist that many of the arguments put forward by connectionism against classical computational approaches are non-demonstrative, I also think that the current way of framing the debate between classicism and connectionism is deeply flawed and bound to give rise to *a priori* and misleading arguments. It is part of the task of this chapter to suggest an alternative way of tackling some of the substantive issues that are at stake between classical and connectionist approaches to cognition.

The two salient claims on which I will focus are that: (i) connectionist modelling can yield neurally plausible models of specific cognitive processes, and (ii) connectionist concepts and principles help to reformulate, and make more tractable a series of interesting aspects of cognitive phenomena.

6.3.3.1 *Neural plausibility*

One of the primary features of connectionist modelling that has been put forward as an index of its adequacy in the study of cognition and as an advantage over classical computationalism is its purported *neural plausibility*. That is, connectionist tools are taken to afford the construction of models of specific cognitive capacities/processes which are plausible from a neurobiological point of view (i.e., compatible with current knowledge about the structure and organisation of the nervous system). Whilst it is true that most connectionist modellers have traditionally drawn their inspiration from the purported computational properties of neural systems (e.g., McCulloch and Pitts 1943), more recently it has become an important point of controversy whether these brain-like systems are really plausible from a neurobiological point of view at all (e.g., Piccinini and Bahar 2013).

The challenge of neural *implausibility* recently raised against connectionism takes one of the following two forms: (i) connectionist models include properties that are not neurally plausible or (ii) they omit other properties that neural systems appear to have. In response to the first type of challenge, some connectionists have claimed that the core features of artificial neural networks might, in fact, be realised in the neural mechanisms of the brain. For example, although the backward propagation of error across the same connections that carry activation signals is generally viewed as biologically implausible, a number of authors have argued that the difference between activations computed based on standard feedforward connections and those computed using standard return connections can, in fact, be used to derive the crucial error derivatives required by backpropagation learning algorithms (cf. Rumelhart, McClelland, and PDP Research Group 1988). Since these latter principles do not pose the same biological implementation puzzle, connectionists have claimed that backpropagation algorithms are not necessarily implausible from a biological perspective after all.

Another defence strategy adopted by some connectionists consists in pointing out that current ANN models are cognitive-level and not proper neural implementational theories, and that they were never intended as such. That is, they claim that connectionists were never particularly interested in modelling specific patterns of connectivity in the brain, or simulating the differences between various neurotransmitters or the very intricate ways in which excitations to neurones are compounded to determine whether a neurone will actually fire or

not. This sort of strategy is supposed to block both types of objections concerning the neural implausibility of current connectionist models. Since most networks model the behaviour of complex systems at levels no higher than cellular and local networks, some abstraction and simplification in the construction of these models is said to be legitimate. This in turn implies that ANNs are abstract models and as such they should not be directly assessed for their neurobiological plausibility. Instead they are intended to capture only certain aspects of the coarse architecture of the brain. Some authors have suggested that this should be regarded as a strength rather than a weakness of connectionism, because it reinforces its status as a proper cognitive theory.

Hence, there seem to be two different stances that one might take with respect to the neural plausibility issue. On the one hand, one may grant that connectionist modelling is underdetermined by the lack of direct neurological relevance of (at least some of) the posits of ANN models (which turn out to be only weakly equivalent with the complex systems being modelled). On the other hand, one may reject the relevance of this type of objection altogether, by invoking the abstract or cognitive character of current connectionist theories of cognition. Either way, the moral of this discussion seems to be the same, namely that connectionism would be better off if it avoided claiming the neural plausibility of its modelling assumptions and posits as a definite advantage over symbolic computationalism.

The purported neural plausibility of connectionist theories of cognition is often invoked in debates between classical computationalism and connectionism in order to show that the latter is an adequate approach for the study of cognition in biological organisms. However, in light of the previous remarks, I contend that, as with other instances of the realisation problem, the issue of the neural plausibility of connectionist models is a red herring. This is not to say that either classical or connectionist models are absolved from attempting to accommodate implementational (realisation) concerns, by proposing, when possible, 'linking-ready' theories/models that could be integrated with neurobiological hypotheses concerning the functioning of the brain. Rather, the point is merely to recognise that, at present, such considerations are too weak to decide between classical and connectionist models of cognition. Next, I turn to another salient feature of connectionism which pertains more directly to the problem of cognitive explanation.

6.3.3.2 *Methodological implications*

One of the signature features of the connectionist approach to cognition is that it promotes a predominant *bottom-up* research strategy which has been taken to have a number of important methodological implications. For instance, connectionists insist that since the focus in

connectionist modelling falls on determining the learning algorithm which is responsible for yielding a certain stable state in a particular type of neural network, one is entitled to substitute, as a default research strategy, experiments in learnability for preliminary theoretical analyses of the domain of interest. That is, given that a function may be learned by a neural network independently of whether we already have a comprehensive theory (or understanding) of the function or not, connectionists tend to claim that it is not necessary to spend a lot of time in analysing the target function before setting a particular network to learn it. This methodological commitment differs strikingly from classical computationalist approaches to cognition where one begins with a characterisation of the adult cognitive capacity (e.g., grammar) and works *backward* to account for how the child might arrive at this 'steady state' characterisation. In contrast, connectionist approaches can be viewed as working *forward* from the existing data about children's cognitive (linguistic) behaviour to some characterisation of adult (linguistic) capacity. Thus, as indicated above, connectionist theories of the 'steady state', in so far as they are available, will tend to be quite unlike theories of such steady capacities put forward by classical theorists.

As a result, the more ambitious underlying claim seems to be that connectionist methods afford a new type of strategy for investigating cognitive phenomena. This strategy is not strictly speaking decompositional or constitutive in the classical sense. Rather, connectionists maintain that artificial neural networks explain higher-level cognitive patterns or regularities by showing that they are the *emergent* consequences of a large number of simple non-cognitive processes. Moreover, whilst connectionism is better characterised as being a piecemeal approach, the connectionist explanatory strategy is said to apply to a (very) wide range (if not all) of cognitive phenomena. Although the question of the precise notion of emergence that is presupposed by these connectionist claims is a matter of some debate (e.g., Stephan 2006; Bechtel and Richardson 1993/2010; Mitchell 2003, 2012), I think that, in light of what has been said so far, one can safely infer that connectionist explanation is a version of the statistical model of scientific explanation.

In fact, neural networks can easily be viewed as an integrative part of a much more general approach to cognition which can be labeled the *generalised statistical modelling* approach (cf. Yang 2002). For instance, the sort of connectionist accounts of linguistic phenomena exemplified in this chapter are part of a more general approach which treats the (child) learner as a generalised data processor that approximates the adult language based on the statistical distribution of the input data. I emphasise this point because connectionists have often drawn on the success of similar approaches, such as dynamic systems accounts, in order to strengthen the generality of their modelling ap-

proach to cognition. But, as shown in chapter 2, characterising something as a statistical explanation does not settle the question of the scope of this particular model of scientific explanation. In particular, viewing connectionist explanations as a form of statistical explanation still leaves open the issue of whether connectionist theories provide the best (and unique) framework for studying cognitive phenomena.

One straightforward problem with this bold implication of the connectionist program is that current connectionist models, for all their virtues, have not yet tackled any interesting cognitive problem in a broad (enough) empirical context. As demonstrated in relation to some of the showcases of connectionist modelling analysed in sections 3.1 and 3.2, the landscape of connectionist modelling is still very fragmented. Hence, connectionists seem to face the challenge of providing first a more unified theoretical account of at least some of the cognitive problems currently under investigation, e.g., object recognition or linguistic inflection. Until that challenge is met, the framework seems to be rather limited with respect to the range of theoretical explanations that it can make available.

In addition, despite the fact that connectionist modellers often denounce the inadequacy of classical theoretical posits, they seem nevertheless to rely on many of the steady-state descriptions provided by classical theories in order to evaluate when and whether a specific neural network has learned a particular set of data and has generalised correctly on the basis of the available input. It therefore seems that theoretical descriptions of cognitive 'steady-states' are implicit in the design of many 'successful' connectionist networks (e.g., they guide the featural analysis underlying distributed representations).

Furthermore, it has been pointed out that, if these theoretical descriptions are not being used, it is much more difficult to establish the explanatory contribution of a particular connectionist model to understanding a given cognitive phenomenon. In such cases, connectionist systems are, at best, viewed as having an exploratory function, i.e., they are used to discover new patterns in certain types of cognitive phenomena. However, the underlying worry here is that the patterns generated by the functioning of a particular network might be mere artefacts of the performance of that connectionist network. Then the justification of the explanatory value of those connectionist models risks to be *ad-hoc* and misleading. This fact implies that the potentially radical character of the methodological implications of the connectionist strategy undermines the very idea that connectionist models are explanatory *at all*.

6.4 CONNECTIONISM: LIMITS AND PERSPECTIVES

In the remainder of this chapter, I propose to synthesise the strengths and the weaknesses of connectionist approaches to cognition. I begin

by reviewing the general lessons drawn from the two case studies discussed in sections 3.1 and 3.2, and show that they do not support the idea of a radical connectionist turn in the study of cognition (4.1). Drawing both on the preliminary conclusions sketched in section 2, and on the lessons afforded by the practice-based perspective adopted in section 3, I then sketch an account of connectionist explanations of cognition (4.2).

6.4.1 *Outcomes of the arguments*

Connectionist approaches have more recently gained increasing influence in various areas of cognitive modelling. In their most radical moments, connectionists claim that PDP architectures provide the foundation for a new paradigm of computational theories of cognition. My investigation has focused on the application of neural networks to specific aspects of language acquisition and development, mainly because it is in this area of research that the clash between connectionism and classicism (i.e., generative linguistics) has been said to be particularly salient.

I have argued that, despite the enthusiasm engendered by the performance of these and other similar networks developed within the cognitive modelling practice, connectionist approaches still face serious problems in matching up to the relevant constraints introduced by both child language acquisition and comparative studies of the world's languages. In addition, given that child and adult language display significant disparities in statistical distributions, adequate connectionist models of language acquisition and development should be able to find a 'pure' learning-theoretic way to account for these biases and for the relation between them (cf. Yang 2002, 2004). As long as these and similar challenges remain unmet, the claim that PDP learning principles/paradigms overturn our (theoretical) conceptions of cognitive capacities such as language is simply overblown.

The critical analysis of the psycholinguistic models carried out in section 3 also encourages the adoption of a more cautious attitude towards the theoretical transformations promoted by some connectionist modellers. The latter claim that connectionist notions and principles impose a radical shift in the theoretical concepts and tools used to describe, analyse, and explain cognitive processes and functions (e.g., language production and comprehension). Rather than endorsing such a radical shift, one should simply acknowledge that connectionist principles sometimes allow us to capture certain features/patterns in the phenomena being investigated that would not be tractable within a different framework (e.g., catastrophic effects in fast learning/memorising tasks, certain types of motion related illusions, etc.), thus increasing our understanding of certain cognitive phenomena. However, recognising the important roles played by con-

nectionist modelling in cognitive research need not amount to rejecting the explanatory role of other theoretical frameworks developed for the study of cognitive phenomena at different levels of analysis and/or abstraction. In particular, it seems premature to say that the limited performances of connectionist networks *refute* the explanatory value of classical notions such as those of compositionality, structure dependency or recursivity.[10]

The preliminary conclusion of the discussion so far is that the strong hypothesis according to which connectionist architectures provide the foundation for a new paradigm of computational theories of cognition is undermined both from a general (theoretical) as well as from a practice-based perspective. Although the principal aim of the proposed argument strategy has been to evaluate critically the arguments and models put forward by connectionists in support of their strong foundational hypothesis, there is also a more positive lesson of this investigation concerning the nature and structure of potential connectionist explanations of cognition, that I will develop in the next section.

6.4.2 *Connectionist explanations*

An important outcome of the critical arguments presented above consists in a moderation of the purported novelty of the connectionist framework. Throughout this chapter, I have striven to show that the emphasis on the novelty of the tools and principles made available on the connectionist framework prevents one from appreciating what a philosophical analysis of connectionism could actually contribute to specific ongoing debates concerning the nature and structure of the mind. A focus on the continuity of the results of connectionist research with existing classical hypotheses seems to be more beneficial than a forced attempt to polarise and compartmentalise the study of cognitive phenomena. Moreover, I claim that the consequence of contesting these novelty claims is an analysis of the roles of connectionism in cognitive research which is actually consistent with the spirit of the integrationist (unificationist) programme promoted by connectionists themselves.

Turning to the problem of connectionist explanation, the philosophical analysis developed in this chapter has revealed three interrelated features that characterise connectionist accounts of cognitive phenomena. Firstly, connectionist accounts are bottom-up approaches on which higher-level (more abstract or general) cognitive patterns or regularities are explained in terms of the interaction and

10 There is also ample evidence (not discussed in this chapter) which suggests that this general call for the 'relativisation' of theoretical notions is in part based on a series of misleading assumptions about the aims and scope of generative linguistics (cf. Wagers and Phillips 2007; Phillips and Lasnik 2002; Phillips 2013a,b).

activity of underlying lower-level (simple/non-cognitive) processing units. However, a distinctive aspect of this bottom-up strategy is that connectionist descriptions and/or explanations are not decompositional and constitutive in the classical sense. That is, the cognitive behaviour or function performed by a certain complex system is not explained, on this sort of framework, by decomposing it into a series of cognitive sub-components which in virtue of their organisation and mutual interaction yield (or are responsible for) the observable higher-level phenomena (as is the case in classical functional analyses). Instead, according to connectionism, cognitive phenomena (e.g., thoughts, utterances, etc.) have a rich complex structure that is the *emergent* consequence of the interplay of simpler (non-cognitive) processes.

Secondly, because it is a bottom-up strategy, connectionism is often said to be consistent with mechanism both on principled as well as pragmatic grounds. On the principled side, connectionism claims that cognitive phenomena are the product of evolutionary and developmental pressures and constraints that include the limited capabilities of biologically realisable hardware and the real-time demands of the environment. In other words, connectionism emphasises the fact that the nature of cognition is shaped by the performance characteristics of the underlying mechanisms, thereby implying that approaches which abstract away from such information run a serious risk of missing critical aspects of the cognitive tasks real biological systems have evolved to perform. On the pragmatic side, connectionists militate for the importance of thinking about specific implementational strategies which in turn may lead to valuable theoretical advances that would be unavailable when operating only at a more abstract level of analysis. For these reasons, artificial neural networks are said to be more adequate tools for modelling specific properties of the neurobiological mechanisms which support cognitive functions/behaviours.

However, the mechanistic undertones of connectionism should not be confused with the adoption of a purely mechanistic explanatory strategy, for, as we have seen in previous chapters, the latter has a strong constitutive and decompositional character which is absent in a proper connectionist strategy. Under a connectionist approach, various aspects of cognitive processing are said to be explainable in terms of the organisation of a class of simple processing units governed by very simple learning rules. This strategy does not seem to rely on any particular assumption about how a complex cognitive process can be decomposed in simpler cognitive sub-processes, or even on any particular localisation hypotheses that would assign specific cognitive functions to particular structures in the nervous system. Thus, I submit that the mechanistic commitments of connectionism are better interpreted as reflecting two salient methodological concerns: (i) the concern to integrate specific implementational hypotheses in a

broader picture of cognition, and (ii) the concern to explain cognitive properties and processes in purely non-cognitive (i.e., mechanistic) terms.

The third major feature of connectionist accounts of cognition becomes evident when one considers the learning functions used to train neural networks to perform particular cognitive tasks. Although I have argued that PDP learnability does not exhaust the analysis of acquisition and developmental phenomena, there are at least two interconnected lessons which follow from the consideration of connectionist learning principles that are too sensible to dismiss. In a nutshell, the intuition strengthened by connectionist modelling is that learning (across a wide range of cognitive domains) is essentially gradual. That is, the sort of statistical learning principles incorporated by neural networks do seem to be most naturally suited to modelling the gradualness of different aspects/parts of cognitive development. But, as argued above, this need not conflict with the hypothesis that different cognitive domains are subserved by specific internal symbolic structures. In other words, the insights made available (both at the theoretical and experimental level of investigation) by connectionism are not necessarily inconsistent with all classical hypotheses concerning the organisation of specific cognitive systems.

Traditionally, connectionist explanations of cognitive phenomena have been characterised as a style of *implementational* or *instantiation* explanations (cf. Cummins 1983, 2010; Fodor and Pylyshyn 1988; Ramsey 1997; Ramsey 2007). Whilst these labels are appropriate in certain respects, they can also be potentially misleading. Given that connectionist models analyse cognitive processes in terms of complex interacting networks of non-cognitive processes, it seems adequate to describe them as implementational accounts/explanations of cognitive phenomena. In addition, at least some ANNs have been used explicitly in the investigation of the neurological mechanisms underlying specific cognitive processes such as episodic memory (e.g., McClelland 1995), which has further encouraged the general conception of connectionism as pertaining to the implementational rather than the more abstract functional or cognitive level of analysis.

The instantiation view of connectionist explanations takes into account the fact that neural networks are quantitatively constrained in a number of ways, viz. with respect to their computing resources, processing time, and the training sets to which they are exposed. Because of these quantitative limitations, neural networks are taken to show how a particular cognitive task would be performed by a very simple system. This is supposed to be the first step of a connectionist explanation which is completed by proving that the functioning of a simple neural network could be scaled up to match realistic constraints concerning both the available biological resources of a human

brain as well as the available input.[11] That is, it has been proposed that connectionist explanations should be characterised in terms of a two-step instantiation strategy, where: (i) the first step proves that a simple connectionist system can perform a specific cognitive task, and (ii) the second step consists in extending the results obtained in the case of a simple network to a realistic connectionist-like system *via* an inductive hypothesis (cf. Matthews 1994, 1997). This way of thinking about connectionist explanations is in line with the implementational characterisation since they both purport to show that (and how) cognitive phenomena are the result of the interplay of a series of simple non-cognitive processing elements.

I agree that one should think about connectionist accounts in implementational terms for the reasons stated above. However, this does not suffice to clarify the sense in which connectionist models of specific cognitive properties and/or processes might be explanatory in the first place. In order to do this, one should further admit that connectionist modelling tools are able to capture certain salient (counterfactual supporting) regularities that would otherwise be missed on a different type of approach (computational or non-computational). And that, in addition, connectionist notions and principles are able to show why and how these regularities occur at all in the sort of (biological) system being modelled. Therefore, I claim that connectionist models play an explanatory role when they generate novel ways to conceptualise and explore particular aspects of cognitive processing. It has been suggested that such a contribution is already available in the form of the statistical learning algorithms investigated by connectionist modellers with respect to different areas of cognitive development.

Rather than insisting that connectionism provides a radical new framework for understanding cognitive processes and their properties, it seems more reasonable to talk about the distinct, and in some cases, complementary contributions that the two computational styles of modelling make to our understanding of cognition. This proposal receives further support from a practice-based perspective which reveals a number of 'hybrid' cognitive models/hypotheses that combine both classical and connectionist principles and hypotheses. For instance, Yang's (2002) variational approach to language acquisition models the process of learning language into a classical generative framework that also incorporates statistical learning principles. The model postulates that learning language is the result of a competition between a number of internally represented (and putatively innate)

11 Thus, most connectionists acknowledge that scalability is necessary for any model to be cognitively plausible (i.e., functional in a realistic world). When applying connectionist models to domains of real-world size and complexity, two problems of scalability can arise: (i) the size of the networks required to implement the modelled capacity may grow out of bounds, and (ii) the time required for the network to learn the required connection weights may become unrealistically long.

grammars (which constitute the hypothesis space) in accordance to a specific statistical learning mechanism. Other examples of hybrid models can be found throughout cognitive science from domains such as vision studies, motor control, up to reasoning, decision making, and language processing theories (e.g., Sun 2008; Marcus 2013). One of the main lessons afforded by this type of models is that unification need not be the definitional goal of a particular framework (such as connectionism), and that instead it should be viewed as the consequence of other more local goals pertaining to the increasing understanding of the nature and structure of cognition.

In contrast, we saw that the adoption of some of the more radical versions of connectionism generates a number of difficult problems. Consider for instance one of the most questionable consequences of radical connectionism. If one accepts that cognitive systems are 'pure' connectionist networks (or generalised statistical learners) whose complex structured higher-order organisation is extracted exclusively from the external environments in which they are embedded, then two things seem to follow. First, every observable cognitive pattern will be explainable in purely statistical terms. Second, and closely related, the connectionist approach would imply that there are no aspects of human cognition that could possibly elude this style of explanation.[12] Taken together, these two consequences seem to preempt the very efforts of characterising the structure of connectionist explanations of cognitive phenomena. Rather than endorsing this strategy, I have proposed that the specific statistical tools deployed in connectionist modelling allow one to capture and explain specific features of cognition (sometimes exceptional or unexpected). Moreover, I have claimed that this proposal is consistent with the idea there are other features of cognitive processing which are more adequately characterised and explained by using other alternative tools or concepts, such as the ones made available by classical computationalist approaches to cognition.

Finally, one might object that the picture presented above becomes possible only because I have deflated too much the opposition between classical and connectionist approaches to cognition. That is, a sceptic might argue that there are substantial differences between the two frameworks at the level of their foundational assumptions and that these disagreements cannot be ignored in assessing the relationship between the modelling practices of the two research communities. In response, I have proposed a strategy that mitigates these

12 It should also be pointed out that from this radical perspective on connectionist explanation, the question of how connectionist systems are individuated simply becomes mute. For if performing any cognitive task depends on the structure of the environment in which the network is embedded (and trained), then the features that distinguish different networks will simply be the ones which are exhibited by the environments themselves and not by the internal constitution and organisation of the network.

foundational issues and provides a critical assessment of the current computational modelling tools employed in different disciplines of cognitive science and neuroscience.

My rationale for preferring this moderate deflationary strategy to a more polemical discussion is that it promotes a way of thinking about cognitive modelling which reveals the continuities between the different existing approaches to cognition as well as some of the general features shared by all styles of cognitive explanations. I claim that, despite its limitations, the proposed strategy manages to avoid a number of spurious debates concerning the meaning of certain theoretical terms used in the cognitive literature. The resulting perspective is one that takes seriously the variety of the explanatory strategies used to tackle cognitive phenomena and tries to account for their individual contributions to advancing our understanding of cognition.

7

A PLURALIST ACCOUNT OF COGNITIVE EXPLANATION

7.1 INTRODUCTION

The main objective of this thesis has been to clarify the nature and structure of explanations of cognitive phenomena. The argument strategy proposed to address this issue combines two complementary perspectives. I began this study with a survey of some of the most influential philosophical accounts of scientific explanation, raising the question whether any of them might provide an appropriate framework for thinking about the nature and structure of cognitive explanations. More precisely, I explored the extent to which these conceptions have been influential in the development of philosophical models of the notion of cognitive explanation. The outcome of these preliminary investigations showed that whilst traditional accounts afford important insights into the structure of scientific explanation, they nevertheless raise a series of issues which bring into question their unconditional application to the cognitive domain. Most importantly, perhaps, traditional accounts of scientific explanation seem to be committed to the thesis of *explanatory monism*, according to which all the sciences conform to a single standard of explanatory 'goodness'. Taken at face value, this seems to challenge the diversity of explanatory schemas actually used in the investigation of cognitive phenomena.

In response, I have argued for the adoption of a more fine-grained, practice-based perspective on the explanatory schemas currently used within the domain of cognitive science itself. Therefore, the second component of the argument strategy developed in the thesis relies essentially on a detailed analysis of several paradigmatic explanatory frameworks used in a number of representative fields of cognitive scientific research. These critical analyses provide a series of ingredients for constructing a more adequate conception of cognitive explanation, that could arguably be extended to other areas of scientific inquiry as well.

This chapter is divided into three distinct parts. Section 2 revisits the principal conclusions of the arguments developed in the previous five chapters of the thesis, exploring some additional critical links be-

tween them. Then, section 3 articulates a *pluralist conception of cognitive explanation* that combines the insights afforded by both traditional philosophical models of scientific explanation and specific accounts of the notion of cognitive explanation. Finally, section 4 discusses the strengths and limitations of the proposed model of cognitive explanation, and considers the prospects of extending it to other scientific domains.

7.2 ARGUMENTS AND CONSEQUENCES

As illustrated throughout the thesis, the difficulty of formulating a substantive account of cognitive explanation arises from the fact that the perspectives identified above seem to impose two conflicting requirements on the project. Traditionally, philosophical accounts have attempted to formulate a highly general and uniform conception of scientific explanation which would hold across time and apply to all fields of scientific investigation. As such, traditional accounts can be taken to be driven by an important *normative concern*, according to which any philosophical analysis of the notion of scientific explanation should capture what distinguishes explanation from other sorts of scientific achievements, by providing *normative* criteria for something to count as a scientific explanation in the first place.

The variety of explanatory strategies that are currently being used in particular domains of scientific inquiry, on the other hand, seems to raise the opposite sort of challenge, namely to provide an account which does justice to the multiplicity and diversity of explanatory tools used by practicing scientists. Thus, according to the latter viewpoint, an adequate philosophical analysis of the notion of scientific explanation must also seek to satisfy a general *descriptive adequacy condition* which requires accommodating and justifying the plurality of explanatory structures/schemas utilised in any particular field of inquiry, such as cognitive science.

Applied to the specific case of cognitive science, the two opposing aims or criteria for a robust substantive account of cognitive explanation are: (i) to characterise in abstract and general terms what makes a cognitive theory/model explanatory *at all* (the *normative* criterion), and (ii) to capture and justify the plurality of explanatory strategies currently used in cognitive science (the *descriptive adequacy* criterion).

Despite this *prima facie* tension, I have insisted that the two perspectives must coexist in order to reinforce and correct one another whenever one of them yields unwarranted claims about the nature and structure of cognitive explanation. I have deployed this twofold argument strategy in chapters 3 to 6, focusing on four influential philosophical models of cognitive explanation. Each of these chapters consisted of two parts. The theoretical part explored the conceptual considerations put forward in support of each particular view

of cognitive explanation and their purported advantages over alternative accounts. The practice-based part assessed each of the proposed philosophical views in light of the modelling and explanatory practices utilised in the relevant field of cognitive science (e.g., neurobiology, computational psychology, psycholinguistics). Whilst the primary outcomes of this study were mainly negative (limiting results), I have argued that they can also be taken to constitute important constraints on a substantive philosophical model of cognitive explanation. In other words, by criticising prominent philosophical models of cognitive explanation, I have also attempted to identify a series of factors which would help to construct a more adequate way of thinking about the problem of scientific explanation in the context of cognitive scientific research.

7.2.1 *Classical models of scientific explanation: Insights and Issues*

The starting point of the present investigations was the critical analysis of three of the most influential models of scientific explanation. The key idea behind this strategy is that, despite the features that make cognitive science a special area of scientific inquiry, there are also numerous methodological aspects it has in common with other natural sciences. Thus, rather than treating the question of what constitutes a proper cognitive explanation in isolation, as a separate and special case within philosophy of psychology or philosophy of mind, I framed this issue as a natural extension of the more general philosophical project of analysing the structure of *scientific explanation*. In this way, I have claimed, both sides can take better advantage of the insights made available by the alternative perspective. That is, debates concerning the nature and structure of cognitive explanation can draw on ideas developed in connection with the broader theme of scientific explanation, while disputes concerning the latter subject matter can build on the hypotheses developed with respect to the special case of cognitive explanation.

The three classical models of scientific explanation surveyed in chapter 2 were: (i) the covering-law model (Hempel and Oppenheim 1948; Hempel 1965), (ii) the statistical/probabilistic model (Jeffrey 1969; Salmon 1971; Mellor 1976), and (iii) the causal model (Salmon 1989; Woodward 2003; Craver 2007b). Each account puts forward a specific category that is supposed to ground the explanatory value of any proposed scientific theory. These are the categories of: (i) natural law, (ii) probabilistic/statistical correlation, and (iii) cause. In contrast to the traditional accounts, I have argued that all these different concepts play a genuine explanatory role if the right kind of relevance relation can be established between the phenomenon circumscribed by the explanandum and the proposed explanatory structure (or explanans). Moreover, the same explanandum can be connected to dif-

ferent explanatory structures that sometimes go beyond those identified by the concepts of law, statistical correlation, and cause. In other words, I have argued that if one pays attention to the particular problems addressed by different scientific explanations and the complex ways in which explanatory relevance relations are established, one is led to conclude that the different conceptual structures identified by each of the classical models typically play an explanatory role only jointly with other categories. This idea mirrors the practice-based observation that often in science explanatory prowess is gained by combining various explanatory schemas that together afford a better understanding of the phenomena being investigated.

These general considerations imply that one should focus primarily on the notion of *explanatory structure* rather than on the idea that explanation should provide a *true account* of the phenomena being investigated. In connection to this, it is important to stress that by focusing on the notions of evidence, reliability, confirmation or truth, mainstream philosophical analyses have tended to neglect a series of factors which play an essential role in the construction of (good) scientific explanations. For although the process of seeking and constructing explanations is inextricably connected with the search for truth, truth by itself is not explanatory. Unlike true statements *simpliciter*, explanations provide insight and enable one's intellectual grasp (or understanding) of certain things treated as being problematic. The emphasis on explanatory structure and the ensuing change of focus from the notion of truth to that of understanding and insight represent the primary result of these preliminary analyses of classical models of scientific explanation. I maintain that this way of rethinking the problem of scientific explanation is especially pertinent if one has in mind the continuity between classical models of scientific explanation and more recent accounts of cognitive explanation. In other words, I take the conclusions derived with respect to classical models of scientific explanation to constitute the general frame of a correct approach to the problem of explanation that can be productively applied to guide the analysis of distinctively cognitive models of explanation.

7.2.2 *The mechanistic view of cognitive explanation*

The mechanistic view characterises cognitive explanation in terms of a *decompositional* analysis that reveals the mechanisms which underlie, support or otherwise maintain the cognitive phenomena being investigated (e.g., Bechtel and Richardson 1993/2010; Craver 2007b; Bechtel 2008). Moreover, mechanistic decompositions are taken to elucidate how a particular phenomenon fits into the causal structure of the world. In fact, the decompositional strategy associated with mechanism underlies both the *causal* and the *constitutive* character of mechanistic explanation. The key presupposition is that the complex

behaviour or function performed by a particular cognitive system is the result of the organised interaction of certain component parts of the system, their properties, and activities. Mechanistic explanations are taken to show: (i) *that* a given system has a particular cognitive capacity in virtue of its component parts, their activities, organisation, and interactions (thus providing a constitutive account of the cognitive capacity); (ii) *why* a particular cognitive phenomenon happened in certain circumstances given its internal organisation and constitution, and (iii) *how* the occurrence of a particular (type of) phenomenon fits in the wider causal structure of the world.[1]

The primary appeal of extending mechanism as an explanatory strategy for the study of cognitive phenomena resides in the promise that mechanistic decompositions will provide an appropriate way of bridging the gap between higher-order characterisations of various cognitive processes and lower-level descriptions of the functioning of certain parts of the nervous system. This aspect of the mechanistic account is important because it challenges the traditional *autonomy thesis* (Fodor 1974; Marr 1982; Fodor 1997), according to which what is relevant for the high-level (psychological) descriptions and explanations of cognitive phenomena is quite different from the kind of thing that is explanatorily relevant from a lower-level (e.g., neurobiological) perspective. Moreover, given the prevalent concern for achieving a more unified theory of cognition, the promise of mechanism seems to be almost irresistible (cf. Craver and Piccinini 2011).

I have argued that both the strengths, as well as the limitations, of the mechanistic view of cognitive explanation are best illustrated by considering some of the paradigmatic examples that have motivated the adoption of the view within the domain of cognitive science in the first place. One example is the neurobiological mechanism that is currently taken to underlie much of higher-level learning and memory processes, namely the mechanism of long term potentiation (LTP), which consists essentially in the long-lasting enhancement in signal transmission between two neurones resulting from their simultaneous stimulation. One of the first models developed in the study of LTP, viz. the Hodgkin and Huxley (1952) equations of synaptic transmission has been used by a number of mechanists for the purposes of clarifying the distinctive explanatory contribution of mechanistic models used in neuroscience (e.g., Craver 2006a; Craver 2007b; Bogen 2008).

The original HH model (Hodgkin and Huxley 1952) described the electrical behaviour of giant squid axon preparations in the form of a mathematical equation which in turn was taken to constitute a first

1 Although the causal and constitutive character of mechanistic descriptions are often conflated in the mechanistic literature (e.g., Craver 2007a; Craver 2007b), it has been pointed out that by differentiating them, one is in a better position to appreciate the variety of problems addressed by mechanistic explanations in the cognitive domain and beyond (cf. Ylikoski 2013).

approximate model of the propagation of action potential. Whilst Hodgkin and Huxley (1952) used fundamental laws (e.g., Coulomb's law and Ohm's law) to derive their equations, by constraining the calculation of ion permeabilities, conductances, resting potentials, driving forces and so on, from experimental data, mechanists claim that the HH equations *do not explain* the transmission of action potentials. Instead, they point out that, from the very beginning, the aim of developing such mathematical convenient characterisations was to facilitate the construction of more specific mechanistic hypotheses/models that show how the factors involved in neural signal transmission yield the observable patterns or regularities. Otherwise put, mechanists argue that although physical laws apply to ion currents and membrane potentials, they do not account for how these factors conspire to produce action potentials in the first place. What is required are proper mechanistic hypotheses concerning the components, properties, activities, and overall organisation of the mechanisms involved in generating action potentials.

Despite their non-explanatory character, mechanists recognise that mathematical characterisations such as the original HH model have been very influential in the development of subsequent mechanistic models of the neurobiological processes underlying learning and memory (e.g., Hille 2001; Naundorf, Wolf, and Volgushev 2006; Ermentrout and Terman 2010). A very similar lesson is derived from the analysis of the difference-of-Gaussians (DOG) model of the space receptive fields of ganglion cells (Rodieck 1965), analysed in detail in chapter 3.

There are three important general lessons that I take to follow from these mechanistic analyses of modelling examples from the field of neuroscience. First, mechanists imply that fundamental laws cannot play the role of proper explanatory structures in the context of a neuroscientific investigation. As such, they oppose the idea that the covering law model of explanation provides an appropriate framework to think about the explanatory practices encountered in the field of neuroscience and/or cognitive science (e.g., Craver 2007b; Bogen 2008). Second, although they acknowledge that abstract mathematical models often play important epistemic roles in the investigation of particular cognitive and/or neurological phenomena, mechanists tend to dismiss the idea that they are genuinely explanatory (Kaplan and Craver 2011; Craver 2012). And, third, what is required for a model to be *genuinely* explanatory is to describe the actual mechanisms that underlie the particular (type of) phenomenon being investigated. In a nutshell, the proposal is that one arrives at such explanatory models via a decompositional strategy that is guided by a series of appropriate experimental (i.e., localisation and control) strategies. Because it assumes that there is a consistent connection between the experimental tools and techniques used by practicing neuroscientists, and

the entities postulated by a mechanistic explanation, the mechanistic conception of cognitive explanation is said to wear its ontological commitments on its sleeve.

Besides its *biological realism*, mechanism has also been promoted for its ability to cover a wide range of models/theories developed in different branches of cognitive science and neuroscience. That is because the decompositional strategy associated with mechanism can be used to obtain a series of mechanistic descriptions of any given cognitive capacity or process ranging from abstract mechanistic sketches, mechanistic schemata, and how-possibly mechanisms up to how-actually mechanistic models (Craver 2007b). Whilst all these mechanistic decompositions can be taken to play a host of roles in the investigation of cognitive phenomena, not all of them qualify as *genuine* explanations. In fact, only how-actually models count as proper mechanistic explanations because they exhibit the *real* causal mechanisms (their components, properties, activities, and modes of organisation) that support or maintain the phenomena under investigation. According to a recent formulation of the *mechanistic explanatory relevance criterion*, called the 3M constraint, a mechanistic model counts as being genuinely explanatory if and only if there is a direct 'mapping' between its component entities and activities, and the actual biological mechanisms underlying the target cognitive phenomenon (cf. Kaplan and Craver 2011).

I have identified three general strategies that might be used to defend such a criterion of mechanistic explanatory relevance, namely: (i) the strong realist strategy, (ii) the moderate realist strategy, and (iii) the epistemic strategy. The *strong realist* strategy has been found to be problematic because it equates the explanatory value of particular mechanistic models with the commitment to the *existence* of certain uniquely determined causal mechanisms (e.g., Strevens 2008; Craver 2012). However, the strong realist claim that there must exist some 'real' mechanisms that explain the target cognitive phenomena leaves open the question whether any proposed mechanistic model counts as being genuinely explanatory. It is of little comfort that we can have *in principle* mechanistic explanatory models of everything, if we do not have a determinate criterion for establishing whether the models that we *do* have are explanatory or not. More problematically, by assuming that everything can *in principle* be explained in a mechanistic way one runs the risk of trivialising the very idea of mechanistic explanation (cf. Psillos 2011).

The most compelling strategy of defending mechanism has been shown to be the moderate realist position. More precisely, I have proposed an interpretation of the moderate realist strategy which mitigates the differences between strongly *ontic* (realist) and *epistemic* views of mechanistic explanation (e.g., Bechtel 2008; Bechtel 2011; Wright 2012). According to the latter type of account, good mech-

anistic explanations are the result of a series of epistemic activities, governed by a variety of norms which connect the evaluation of the explanatory value of particular mechanistic models to the epistemic aims and interests pursued by different research communities. In some of its formulations, the epistemic view has been taken to entail a stronger relativist position according to which any proposed mechanistic model can be deemed to be explanatory relative to some set of aims and purposes. I have argued that one way to avoid this sort of trivialisation of the mechanistic conception is to recognise that both epistemic and ontological principles play a role in the development of good mechanistic explanations of cognitive phenomena (cf. Illari 2013).

This proposal shifts the emphasis from the question concerning the 'source' or 'locus' of the explanatory value of particular mechanistic models towards the norms that 'locally' govern the practice of developing mechanistic explanations in the cognitive domain. I have argued that this perspective allows one better to appreciate the difference between the *explanatory structure* of mechanistic models and the *norms* which govern the construction and evaluation of explanatory models of specific cognitive phenomena. The explanatory structure of mechanism consists in a decompositional analysis of the target system into component parts, their properties, and activities, whose organised interaction can be taken to yield the phenomena or pattern being investigated. Viewed as a special type of explanatory structure, mechanisms can be differentiated both from laws and causal links *simpliciter*.

This way of looking at mechanism has two welcome consequences. First, if mechanism is no longer viewed as an absolute explanatory category then it becomes more straightforward, identifying the various epistemic roles played by mechanistic decompositions and showing how they may enter an explanatory strategy together with other components, such as mathematical concepts. For instance, I have claimed that the Difference of Gaussians (DOG) model of the organisation of the spatial receptive fields of early visual neural cells (Rodieck 1965) allows one to understand the special selectivity of these neurones to certain formal geometrical features of a large variety of visual stimuli. However, I have also pointed out that the explanatory value of abstract (mathematical) models does not rule out the possibility that these models can also be used as launchpads for the construction and refinement of other mechanistic models of cognitive phenomena. Rather, the lesson derived from these considerations is that mechanisms are better viewed as elements of more complex explanatory schemas that are deployed in order to elucidate particular cognitive phenomena.

Second, the arguments developed in chapter 3 showed that the unificatory potential of the mechanistic framework is conceptually dis-

tinct from the explanatory value of mechanistic models/theories of cognitive phenomena. Thus, against Craver (2007b) and Craver and Piccinini (2011), I have argued that the unificatory power of the mechanistic framework does not directly entail mechanistic explanatory monism. In fact, conflating the two epistemic virtues risks blurring the mechanistic explanatory relevance criteria which in turn mirror the epistemic and ontological norms/principles that guide the practice of developing and refining mechanistic models of cognitive phenomena.

7.2.3 *Classical computationalist explanations*

Having identified the strengths and limitations of the mechanistic conception, I have then focused on another influential model of cognitive explanation that has been widely discussed both in the philosophical and cognitive scientific literature. In a nutshell, classical computationalism is the view that cognitive capacities and/or processes are explainable in terms of internal mental representations (symbols) and operations (rules) appropriately defined over them. Although classical computationalism also qualifies as an analytic decompositional strategy, accounting for complex cognitive capacities in terms of their component sub-capacities and organisation, the view is taken to differ from mechanism because of its lack of explicit neurobiological (mechanistic) commitments. The latter idea supposedly follows from the *autonomy thesis* standardly associated with classical computationalism (e.g., Fodor 1974; Marr 1982; Cummins 1983; Block 1997; Fodor 1997). On a strong reading of this thesis, computational explanations of specific cognitive phenomena can be developed independently of hypotheses pertaining to the neurobiological (or implementational) level of organisation of cognitive capacities.

The arguments put forward in chapter 4 explored two important aspects of the view of cognitive explanation associated with classical computationalism: (i) the structure of classical computationalist explanations of cognitive capacities, (ii) the criteria for individuating the computational states and/or structures postulated by such explanations. The first issue amounts to clarifying the factors that contribute to the construction of adequate computational explanations of particular cognitive phenomena, whereas the second concerns the criteria that are involved in determining the type-identity of particular computational states/structures. By clarifying the relationship between these aspects of classical computationalism, the chapter sheds light on another topic that has attracted much attention in philosophy of mind and cognitive science: the notion of *mental content* and its function in cognitive theorising. With respect to this problem, I have attempted to develop a position which elucidates the role(s) that mental contents play in the construction and evaluation of potentially

explanatory computational models/theories of cognitive phenomena, whilst remaining neutral with respect to the question of what would constitute a proper philosophical theory of mental content.

My primary interest was to present and analyse classical computationalism as one of the alternative explanatory frameworks utilised by practicing cognitive scientists, rather than as a broadly metaphysical picture of the mind and its place in nature. As a consequence, I have focused on a series of philosophical arguments that seemed to be directly relevant for the central issue of cognitive explanation, treating other philosophical hypotheses that are standardly associated with these different classical computationalist accounts as orthogonal to the main theme of the thesis. The argument strategy adopted for this purpose relies essentially on the systematic distinction between the *computational individuation* issue and the issue concerning the *explanatory value* of particular computational models of cognitive phenomena.

As a starting point, I have identified the core principles shared by the various theoretical accounts which have been developed under the wide umbrella of classical computationalism in both the philosophical and scientific literature (e.g., Fodor 1980; Pylyshyn 1984; Egan 1992, 1999, 2010; Gallistel 1993; Gallistel and King 2009). The key idea that these views share is that certain aspects of cognitive phenomena can be appropriately explained in terms of a set of operations defined over symbolic computational structures or states. This commitment is also captured by the hypothesis that classical cognitive architectures consist in a set of rules and symbols which, when appropriately manipulated or transformed, yield the target features of cognitive phenomena. This general model of cognitive explanation raises two interrelated questions: (i) what makes something a computational system in the first place, (ii) what are the grounds for believing that computational systems are an appropriate tool for investigating and explaining cognitive phenomena. The first question is the target of the *computational individuation* issue, whereas the second question concerns the justification of the *explanatory value* of (classical) computational models/theories of cognition.

With respect to the individuation issue, I have shown that a wide range of arguments advanced by proponents of classical computationalism entail an *internalist* or *formal* view of computational individuation (e.g., Fodor 1980; Chomsky 1995; Egan 1992, 2010, 2013). According to such an internalist thesis, the states and structures of a computational system are individuated in virtue of their formal or abstract properties. On this view, the computational structure of a system coincides with its formal (or syntactic) structure, which it implements. The internalist view of computational individuation opposes the idea that content impacts the type-identity of a computational system. In these debates, perhaps the most widespread notion

of content invoked (but by no means the only one) is that of *external* content, by which one typically means some feature or other of the system's external environment. Thus, according to what has been called the *semantic view of computational individuation*, which opposes an internalist account, semantic features (such as factors in the system's external environment) determine (in part) the computational identity of certain physical systems (e.g., Burge 1979, 1986; Wilson 1994; Shagrir 2001).

The hypothesis that content impacts computational individuation is strongly connected to the idea that the component structures/states of computational systems (i.e., symbols) have a dual nature. That is, they possess purely formal or syntactic features and they can be assigned semantic interpretations (contents). I have surveyed a number of arguments (e.g., Fodor 1980; Stich 1983, 1991; Egan 1992, 1999, 2010; Matthews 2007) which purport to show that symbols do not possess their contents essentially and that content - either *broad* (externalist) or *narrow* (i.e., narrow causal roles or formal features such as high-level mathematical relations among the represented objects) - does not determine the computational type of a particular physical system. I have shown that some of these arguments draw on the mismatch between the stability of computational taxonomies and the context-sensitivity and ambiguity of semantic (intentional) characterisations of cognitive capacities (Stich 1983, 1991; Egan 1999), whereas others insist that the semantic view of computational individuation does not reflect the actual theorising and modelling practices of computationalist cognitive scientists and neuroscientists (Egan 1992, 2010; Chomsky 1995).

Whilst I have taken these arguments to be successful in showing that the type-identity of a particular computational system depends only on its component entities, their internal relationships, and organisation within the system, I diverge from these positions when they take this internalist view of computational individuation to entail a purely internalist picture of (classical) computationalist theories of cognitive capacities. My primary motivation for defending an internalist or formal view of computational individuation is that such a view allows one to understand computational systems as independent epistemic tools which can be used in the investigation of cognitive phenomena. That is, they can be characterised as having certain properties which make them appropriate for the explanation of some aspects of cognitive phenomena. I have claimed that this view of computational individuation is neutral with respect to the metaphysical hypothesis that the mind has an essentially classical computational structure. In addition, the proposed internalist individuation hypothesis is consistent with the idea that computational theories of cognition can and sometimes do make substantial reference to features of the external environment of the target cognitive system.

Nevertheless, I have claimed that in order to support this weak form of externalism with respect to computational theories/models of cognition, one does not need to be committed to the idea that content impacts computational individuation *per se*. Instead, I have argued that even more sophisticated accounts such as those proposed by Shagrir (2001) and Egan (1999, 2010) are best interpreted as saying that contents play an important role in the *identification* of the specific type of computational system which explains a given cognitive capacity. For instance, Shagrir (2001) argues that in order to be able to determine which syntactic (formal) characterisation of a computational system is the most appropriate with respect to a given cognitive problem/task, one needs to take into account some of the semantic features of the computational system in question (e.g., formal features such as set-theoretic or other high-level mathematical relations among the represented objects). Similarly, Egan (1999, 2010) contends that mathematical functions provide canonical descriptions of the functions computed by particular types of physical systems. However, there are good reasons to doubt that either of these positions manages to show that a sufficiently substantive notion of content determines the type-identity of computational systems.[2]

Instead, I have taken Shagrir's (2001) observation that 'there is a close relationship between a system's computational identity and the semantic characterisation of the [cognitive] task in question' (ibid., p. 17) to delineate the *explanatory* aim of a computational model of that particular cognitive task. That is, if a computational model of a particular cognitive capacity is to count as genuinely explanatory, one must be in a position to show how the posits of that theory relate to the cognitive problem under investigation. Along these lines, I have claimed that (mental) contents play an essential role in connecting a computational model to the cognitive phenomenon or pattern that it is taken to explain. Otherwise put, in order for a particular computational model to count as an explanation of a specific cognitive pattern one must be able to show how certain elements of the model relate to the features of the explanandum phenomenon which are standardly characterised in semantic or representational terms. As shown in the analysis of the three models from vision studies surveyed in chapter 4, the types of contents assigned to the states of particular computational models can be either broad (externalist) or narrow, depending on the cognitive task the system is taken to solve. Moreover, I

2 I have pointed out that a version of 'internal semantics' would be compatible with the proposed view of computational individuation, but the latter would in turn be characterised solely in terms of the internal relationships and organisation of the elements of the computational system. In other words, internal semantic contents would fail to have truth or correctness conditions which would be definable independently of the internal organisation of the system. Thus, insofar as they impact the computational-type of a given state or structure they would count merely as formal or abstract features of the system in question.

have claimed that this variability of the types of semantic contents assigned to the states of particular computational models is perfectly consistent with the hypothesis that the computational type-identity of those systems depends entirely on the internal constitution and organisation of the system.

In brief, I have argued that the assignment of both broad and narrow (mental) contents is extrinsic to the constitution and functioning of a computational system/model. However, unlike other accounts, the proposed view does not take the extrinsic character of semantic interpretations of classical computational models/theories to entail that computational theories/models of cognition are purely internalist or individualistic. In other words, one cannot derive a complete picture of the explanatory value of classical computational models of cognitive phenomena solely on the basis of the *internalist view of computational individuation*. In fact, doing so would be just another way of conflating the individuation and explanation issues all over again. What is required in addition is a more precise account of how computational systems/models may be successfully applied in the study of cognitive phenomena. In relation to the latter task, I have argued that semantic interpretations (or mental contents) play an essential bridging role in connecting the explananda of computational theories of cognition to the proposed potentially explanatory computational structures.

Moreover, insofar as (some of) the explananda of these computational theories are characterised (sometimes) in broad externalist terms, one can also say that *computational theories of cognition* are externalist. However, from this it does not follow that the type-identity of the computational structures postulated by these theories is determined by any of the proper semantic features of their explananda. Thus, defenders of an internalist view of computational individuation need not reject the idea that semantic interpretations (contents) constitute a *proper* part of computational theories/models of cognition. Whilst they are not constitutive of the proposed computational models, they play a normative role in that they serve to justify why a given computational system, which is formally individuated, plays a proper explanatory role with respect to a particular cognitive phenomenon.[3]

In light of these considerations, I have argued that classical computational explanations of cognitive capacities typically proceed by decomposing a complex cognitive task (e.g., visual object recognition) into a set of more basic cognitive tasks (e.g., edge extraction, feature construction and ordinal matching, etc.) which in turn are characterised in terms of a series of computing operations defined over

3 In addition, I have suggested that calling this justificatory role played by mental contents, the 'informal' part of the computational theory (Chomsky 1995) or the 'pragmatic gloss' (Egan 2010) can distort the analysis of computational models of cognitive phenomena.

appropriately typified symbols. Thus, computational explanations connect, via a number of identifiable steps, the target phenomenon (which is often, but not always, characterised in semantic terms) to a computational structure which reveals certain fundamental features of the cognitive phenomenon in question. These features contribute to a better understanding of cognitive phenomena by allowing cognitive scientists to test a wider range of counterfactual generalisations pertaining to them, and to draw connections between the computational characterisations of apparently disparate cognitive capacities. The assignment of mental (semantic) contents plays a crucial role in the development of these computational models because it serves to justify (at least in part) why a particular computational structure/system can be taken to capture something relevant about the structure of the target cognitive phenomena. I have insisted that in order to play this sort of normative role in the construction and refinement of (good) explanatory models of cognitive phenomena, the interpretation/assignment function need not be taken to be a one-to-one mapping between computational states and semantic contents.

In fact, even a cursory glance at the modelling practices of cognitive scientists shows that semantic interpretations of computational models are most of the time partial and mixed (i.e., comprising both what philosophers have identified as broad and narrow contents). Furthermore, I have pointed out that the explanatory adequacy of particular computational models is usually evaluated by taking into account additional constraints, developed along the lines of the *strong equivalence criterion* (Pylyshyn 1984). According to this, the input-output equivalence of the cognitive system and the computational structure postulated by a particular computational theory/model may not suffice to guarantee that the latter has a genuine explanatory value. In addition, one might need to take into consideration measures such as response times, complexity profiles of the modelling and modelled systems, and so on, which would further increase the adequacy of the computational structures postulated by classical computational models/theories of cognition.

At a more general level, the arguments developed in chapter 4 advocate a shift of focus within the debates concerning classical computationalism from the metaphysical import of the classical computationalist thesis and the prospects of a substantive theory of mental content towards an analysis of the epistemic practices which generate explanatory models of particular features of cognitive phenomena.

7.2.4 *The mechanistic view of computational explanations*

The computational account of cognitive explanation analysed in chapter 5 appeals to the main tenets of the mechanistic view of explanation (e.g., Machamer, Darden, and Craver 2000; Craver 2007b; Bechtel and

Richardson 1993/2010) in order to elucidate the structure and scope of computational models/theories of cognitive capacities. As such, the account promises to provide a better view of computational explanation which is in line with the principles used by practising scientists in constructing and refining computational theories of cognitive capacities. The analysis focused on two main features that have been taken to distinguish the *mechanistic view of computational explanations* from the classical computationalist account. First, the mechanistic account is said to entail a non-semantic view of computational individuation which avoids some of the problems standardly raised for classical computational accounts. Second, according to its advocates, the mechanistic conception provides a principled way of bridging the gap between the higher-order computational descriptions of cognitive capacities/processes and their neurobiological characterisations.

According to one of the most elaborated versions of the *mechanistic conception of computationalism* (Piccinini 2007a, 2008a; Craver and Piccinini 2011), computational explanations are a sub-species of mechanistic explanations. More precisely, the mechanistic view of computationalism comprises two purportedly distinct hypotheses: (i) a *wide functional individuation* hypothesis, and (ii) a *mechanistic explanation* hypothesis. The wide functional individuation hypothesis claims that computational states are individuated by their wide functional properties, which in turn are specified by a mechanistic explanation in a way that need not refer to any semantic properties. As shown in chapter 3, a mechanistic explanation decomposes a given complex system into its component parts, their functions, and organisation, and claims that the system exhibits a particular capacity (or set of capacities) because it is constituted by the relevant components, their functions, and their organisation. Proponents of mechanism have insisted that the wide functional individuation strategy differs in significant respects from the hypothesis that computational states/structures are individuated in terms of their wide functional contents (cf. Piccinini 2008a). The latter, but not the former, counts as a semantic view of computational individuation.

There are two main arguments that have been put forward in support of the idea that wide functional individuation is a non-semantic individuation schema. The first consists in the criticism of the semantic view of computational individuation (cf. ibid.). The latter is said to rely on a systematic confusion between computational and content individuation, imposing a view of computation that is hostage to a series of metaphysical intuitions about the nature and structure of cognitive capacities. This is taken to be problematic because it obscures the fact that computational systems can be characterised independently of a cognitive context. A closely related criticism is that the semantic view of computational individuation ignores the wide range of modelling and theorising practices in which computational

systems are being used and which do not seem to rely on any semantic individuation schema.

The second argument draws on the connection between computational individuation and the mechanistic organisation of computational systems. As suggested above, the positive argument for the wide functional individuation strategy relies on the fact that computational systems are amenable to mechanistic decompositional analyses. That is, mechanists have argued that the computational states of a particular physical system are individuated in terms of the functional properties of the component parts of the system that have been identified by the relevant decompositional mechanistic analysis. Otherwise put, the mechanistic analysis is taken to establish the functionally relevant properties that play a role in the individuation of specific types of computational states. However, the problem with this proposal is that the mechanistic strategy *per se* cannot guarantee that the functional properties it is capable of identifying will be computationally relevant as well. If wide functional individuation is to count as a proper computational individuation schema, mechanists have to show what distinguishes the functional properties that are computationally relevant from those that are not.

I have analysed one proposal that attempts to address this sort of indeterminacy problem which threatens the wide functional view of computational individuation. In a recent paper, Piccinini and Bahar (2013) have argued that mechanistic analyses of computational systems count as computationally relevant only those functional properties that are essentially related to the characterisation of the system in terms of a *generic computational system*. A generic computational system in turn is defined as a mechanism that comprises a series of entities (or variables that can change state) which are transformed or manipulated in accordance with rules that are sensitive only to certain properties of the system's component parts. The notion of generic computation is taken to be the common denominator of a larger class of computing systems, which includes digital computers, analog computers, and neural computing systems (cf. ibid.). In the case of computational models of cognitive capacities, the relevant subspecies of generic computation is that of *neural computation*. Thus, mechanistic analyses of proper neural computational systems will single out only those functional properties which can be attributed to the entities over which neural computations are appropriately defined, namely properties of neural spike trains (e.g., firing rate and spike timing). It is these functional properties which are taken to play a role in the type-individuation of specific computational systems.

This appeal to the notion of neural computation is supposed to fulfil two tasks: (i) to solve the indeterminacy problem by constraining the range of functional properties which are relevant to the computational identity of a given system, and (ii) to demonstrate that it is

possible to circumscribe a neurobiologically plausible notion of computation. Both points have been taken by proponents of the mechanistic view of computational explanation to show the advantages of their position over the classical computationalist view of cognitive explanation (Piccinini 2008b; Piccinini and Bahar 2013). In response, I have argued that the differences between the two views are less dramatic than advertised in the mechanistic literature. With respect to the individuation issue, I have argued that the functional view of computational individuation is perfectly consistent with the internalist conception of classical computationalism, argued for in chapter 4. At most, mechanists can claim that their proposal is less vulnerable to a certain type of ambiguity which derives from the idea that the computational-type of a particular physical state or structure depends on its intrinsic formal properties and its relations with the other similar structures. In addition, mechanistic analyses can be taken to provide a more precise characterisation of these internal properties and relations.

Concerning the problem of computational explanation, mechanists claim that a computational model/theory of a given cognitive capacity counts as being genuinely explanatory if and only if it respects the functional individuation criteria and it is defined over the appropriate types of entities, which they take to be spike trains and some of their properties, such as firing rates and spike timing (ibid.). However, I have pointed out that there are good reasons for resisting the adoption of the notion of neural computation as the only one that can serve as an explanatory structure in the investigation of cognitive phenomena. On the one hand, the notion seems to be limited because it can be used to explain and elucidate cognitive patterns/phenomena only at certain levels of organisation or resolution (e.g., neurobiological levels). That is, given that neural computations are defined over properties of biological neurones, the explananda with which these structures might be appropriately connected do not seem to go beyond the patterns studied in certain branches of neuroscience. Thus, it is not clear how, and indeed whether, the notion of neural computation can actually be used to elucidate cognitive phenomena or patterns characterised at higher-levels of abstraction (e.g., patterns such as long-distance dependencies and island constraints in language production, or certain systematic patterns in reasoning and decision making). More problematically still, there are reasons to doubt that there are yet any such appropriate mechanistic computational models of cognitive phenomena at any level of analysis or abstraction.

On the other hand, if one accepts that not all explanatory accounts of cognitive capacities need reference their precise neurobiological make-up, then it is possible to show that mechanism is compatible with the idea that there are abstract computational characterisations

of certain cognitive phenomena which advance our understanding of them by revealing a host of important features of the phenomena in question. Thus, I have argued for a type of continuity between the classical computationalist and mechanistic view of cognitive explanation which does not imply that either of the accounts should be reduced to the other. Instead, the mechanistic view of computational explanation can be taken to highlight the implicit mechanistic commitments of classical computationalism. In addition, I have suggested that where the mechanistic model diverges from the classical one, it makes available potentially new explanatory structures (neural computations) which can be used to construct additional partial explanations of specific features of cognitive phenomena.

7.2.5 *Connectionist explanations*

Connectionism claims that cognitive phenomena/patterns can be explained in terms of a series of internal computations which are carried out by a set of simple processing units that operate in parallel and affect each other's activation states via a network of weighted connections. Although the core ideas that underlie connectionist approaches to the study of cognitive phenomena can be traced back to the work of Alan Turing (Turing 1948/2004; Copeland and Proudfoot 1996; Teuscher 2002), who also inspired classical computationalist views, connectionism has been standardly construed in opposition with the notion of classical or symbolic computation (Rumelhart, McClelland, and PDP Research Group 1986; Rumelhart, McClelland, and PDP Research Group 1988; Bechtel and Abrahamsen 1991; Elman 1996). A different sort of relationship holds between connectionism and mechanism, with the former claiming to be consistently committed to mechanistic assumptions on both principled and pragmatic grounds (Bechtel and Abrahamsen 1991; Bechtel and Richardson 1993/2010; McClelland et al. 2010). The primary aim of chapter 6 was to identify the main tenets of connectionism and to elucidate the view of cognitive explanation associated with this research strategy.

As in the previous two chapters, I have divided the analysis of connectionism into two distinct strands which reflect the two complementary perspectives developed throughout the thesis: the theoretical and the practice-based perspective. I started by discussing the problem of computational individuation in a connectionist setting. This served not only to get a better grasp of the theoretical principles underlying much of current connectionist modelling but also created a substantive base for drawing a more balanced comparison between classical and connectionist computationalism. With respect to the individuation issue, I have argued that there are two major strategies for determining the type-identity of specific connectionist networks:

(i) an input-output or functional strategy, and (ii) a more fine-grained architectural strategy.

On the functional individuation strategy, a connectionist network is individuated in terms of the input-output function (or pairings) which it computes. Whilst the functional individuation strategy brings connectionism closer to the classical view of computation, it tends to reduce the potential diversity of connectionist networks that can be used in the study of cognition by ignoring a number of features which are standardly taken to constitute the characteristic marks of connectionist networks. The architectural strategy on the other hand, takes into account many more of the intrinsic parameters used to define a connectionist network, and, in consequence, yields a more fine-grained taxonomy of connectionist networks that might be used in the study of particular cognitive phenomena. At a more general level, what the two individuation strategies suggest is that (as with symbolic computational systems) neural networks are complex systems whose functioning and organisation can be understood independently of their applications to the study of specific cognitive phenomena. Also, both individuation strategies are consistent with an internalist point of view, according to which the computational identity of a particular physical system does not depend on any semantic interpretation that can be assigned to it in a particular cognitive modelling context.

I have shown that connectionism purports to offer a non-decompositional style of explanation in which specific cognitive patterns are said to be the *emergent* consequence of the controlled propagation of activation patterns through a network of very simple units. One way that has been proposed to capture this idea is to conceive of connectionist explanations as a two-step explanatory strategy in which one first attempts to show that a rather simple network appropriately organised and trained can yield a particular type of outputs in response to the relevant class of inputs. Then, the second step of a connectionist explanation requires one to extend the positive results observed in the case of the simple network to scaled-up realistic models of the nervous system (cf. Matthews 1997; Ramsey 1997). This two-step model of connectionist explanations is meant to reinforce the idea that connectionist models are not mere implementational variants of their classical (symbolic) counterpart models, but provide a genuine novel way of accounting for a wide range of cognitive phenomena.

By analysing several paradigmatic connectionist models used in the study of language acquisition and processing, I have attempted to show that this two-step explanatory strategy does indeed seem to be presupposed within the connectionist modelling practice. The proposed analysis has served three additional purposes: (i) to draw out the limitations of the particular models under discussion, (ii) to illustrate a range of theoretical claims made by connectionist mod-

ellers and philosophers, and (iii) to pin down the main challenges facing current connectionist approaches to cognition. In particular, I insisted that the radical connectionist position runs into a number of difficult issues which undermine the very possibility of articulating a coherent notion of cognitive explanation. For given the strong polarisation promoted by resolute connectionists between classical and connectionist architectures and/or modelling principles, one is led to believe that connectionist modelling and theorising is not guided by any steady state description of the cognitive system/phenomenon being investigated. This further raises the question whether there is another way in which one might circumscribe the proprietary cognitive explananda of connectionist theories/models.

I have explored one compelling way of characterising the explanatory value of connectionist models of cognitive capacities. Connectionists have insisted that neural networks are an appropriate tool for investigating certain features of learning and development processes which lead to the constitution of the cognitive capacities whose steady-state descriptions are the object of other theoretical approaches such as classical computationalism or mechanism. For instance, connectionist principles have been said to capture and account for the gradual character of language acquisition and language change (cf. Marcus 2001; Yang 2002). In addition, connectionism seems to provide the appropriate framework for explaining singular or unexpected patterns in cognitive processing such as certain types of catastrophic interference phenomena in learning and memory tasks (cf. McClelland et al. 2010). Whilst such contributions of connectionist modelling are indicative of the potential explanatory value of connectionist principles in cognitive theorising, they do not support the stronger contention that connectionism provides the unique explanatory framework appropriate for studying cognitive phenomena at all possible levels of abstraction or analysis.

Moreover, I have pointed out that one objectionable consequence of radical connectionism is that it reverses the order of cognitive explanation, letting connectionist models determine the cognitive phenomenon to be explained. Whilst this aspect of radical connectionism might be taken to reflect the bottom-up character of the connectionist methodology, it also undermines the idea that connectionist networks are epistemic tools that advance our understanding of specific aspects of cognitive phenomena which we deem to be problematic. At best, the strict bottom-up character of connectionist modelling can be taken to indicate that neural networks might be used for investigative purposes, in order to discover new and perhaps unexpected patterns in cognitive processing. In addition, I have pointed out that the actual modelling practices of cognitive scientists show that the applicability of connectionist models to the study of specific cognitive phenomena depends to a significant degree on the semantic interpretations

assigned to specific connectionist models. Whilst 'shallow', i.e., applicable only to the input-output interface of the connectionist models themselves, these interpretations play an ineliminable role in securing that particular neural networks are adequate explanatory models of specific aspects of cognitive phenomena.

In analysing the premises of the radical connectionist position, I have also suggested that the other major putative advantage invoked in support of connectionist theories of cognition concerning their neurobiological plausibility should be taken with a grain of salt. As in the previous chapters of the thesis, I have pointed out that the realisation issue cannot be used to settle definitively the dispute on the explanatory value of specific computationalist models of cognition. However, considerations concerning the biological plausibility of connectionist models/theories of cognitive phenomena have been taken to align connectionist and mechanistic assumptions. From a practice-based perspective, one can clearly point to cases where the connectionist methodology has been used to obtain further evidence and control over the neurobiological mechanisms underlying specific cognitive processes (e.g., McClelland 1995; Eliasmith 2007; Thomas and McClelland 2008; Eliasmith 2010). However, the relationship between the two frameworks seems more problematic when one considers the structure of connectionist and mechanistic explanations, respectively. For, on the one hand, connectionism promotes a non-decompositional explanatory strategy in which one explains a given cognitive phenomenon as the emergent consequence of the activity of a huge set of simple units, whereas, on the other, mechanism relies on a constitutive decompositional strategy.

I propose that the solution to this problem is again the partiality of cognitive explanation. That is, connectionist networks can be used to capture important aspects of cognitive phenomena which advance our understanding of the phenomena in question. In addition, connectionist models sometimes afford a series of insights about the neurobiological organisation of the mechanisms which underlie different cognitive tasks (e.g., memory and learning tasks). Whilst this latter point supports the idea that connectionist principles have multiple applications in current cognitive and neuroscientific research, it does not prove that mechanism and neural networks function as the same type of explanatory structure and can be used in the same explanatory contexts. For instance, I have pointed out that connectionist principles and/or concepts seem to be more appropriate than mechanism to account for highly improbable cognitive effects, in which case the model of connectionist explanation is a sub-species of the more general probabilistic/statistical view of scientific explanation.

In summary, I have claimed that the better policy is to evaluate on a case by case basis the explanatory contributions of individual connectionist models without holding onto a resolute connectionist per-

spective according to which all cognitive phenomena will be someday explainable in a purely connectionist framework.

7.3 COGNITIVE EXPLANATIONS

Drawing on the results of the critical analyses summarised above, in what follows I propose to outline a pluralistic picture of cognitive explanation. The view of explanatory pluralism that I endorse attempts to do justice to the multilevel and multicomponent (interdisciplinary) character of cognitive scientific research. For this purpose, it emphasises and expands three general features of the structure and organisation of cognitive science as a field of interrelated disciplines which investigate, at different levels of analysis and abstraction, various aspects of cognitive phenomena.

First, within each explanatory framework (e.g., mechanism, computationalism, connectionism, etc.), there are a series of norms or principles that guide the construction and evaluation of good models/theories of cognitive phenomena. They comprise both ontological and epistemic constraints which have been crystallised in the experimental and theoretical activities of the relevant scientific community. Second, models/theories of cognitive capacities (from reaching and grasping, and object recognition, to language processing and decision making) can sometimes be in competition with one another (such is the case with models of language processing which postulate different derivational orders or the existence of innate rules or probabilistic learning principles). When such substantial disagreements arise, it becomes necessary to decide which of the proposed accounts constitutes the most adequate model of the cognitive capacity under investigation. Epistemic virtues such as the explanatory value of the competing models, their empirical adequacy, or unificatory power usually play a crucial role in evaluating the relative value of competing cognitive theories or models. Third, on occasion, scientific models/theories developed at different levels of analysis or abstraction can be compatible with one another and, when this happens, they can be (and typically are) coordinated into a more complex account of the target cognitive capacities or processes (e.g., Shadmehr and Wise 2005). In these cases, the explanatory factors which operate at different levels of analysis or abstraction are combined in order to yield a mixed-level account of the phenomenon in question (cf. Mitchell 2003; Wilson 2010).

Thus, one of the major claims I make in this thesis is that different research programmes in cognitive science give rise to corresponding distinct explanatory frameworks, each one of which comprises a complex of specific explanatory structures (concepts) which contribute to advancing our understanding of particular aspects of cognition and its underlying biophysical mechanisms. In other words, I promote

the adoption of a form of *epistemic* explanatory pluralism according to which different frameworks can be said to contribute in distinct ways to increasing our understanding of the problems addressed at different levels of analysis or abstraction in the various branches of cognitive science and neuroscience.

7.3.1 *A pluralist view of cognitive explanation*

In what follows, I propose to spell out the main tenets of a *pluralist conception of cognitive explanation*, building on various insights expressed earlier. However, the proposed account should not be taken as a direct generalisation from the preceding chapters, but rather as an articulation of the abstract framework that has been presupposed by the critical analyses developed so far.

The general model of cognitive explanation that I put forward draws on the intuition that many different conceptual structures can play an explanatory role with respect to different types of (empirical) problems. In order to get this proposal off the ground, I start from the idea that explanations are sought and offered when certain aspects of reality are deemed to be problematic. That is, assuming that certain things are known, other things might be construed as problematic. In this sort of context, explanations are supposed to create ways (strategies) to resolve the tension between what we know and what we take to be problematic. For instance, given our knowledge about the properties and organisation of ganglion cells in the early visual system, and the complexity of the visual input, the question arises how do these cells contribute to the extraction of certain patterns or features of the visual input that are believed to play a role in higher-order cognitive tasks such as object recognition. Or, given what we know about the structure of natural languages and children's early exposure to linguistic environments, how can we account for the gradual character of language acquisition and/or for the various stable patterns that characterise different stages of language development in small children.

One important idea that has emerged from analysing the various explanatory frameworks used in cognitive scientific research is that explanations impose certain types of conceptual structures on their explananda which in turn enable a better intellectual grasp of the phenomena under investigation. Otherwise put, an explanation creates a link between a particular characterisation of a target phenomenon (the explanandum) which construes it as problematic and another description (the explanans) which shows how the phenomenon in question is to be understood within a given system of scientific knowledge. This general picture of explanation does not imply that all requests for explanation are met by appealing to the same well-established collection of concepts. Sometimes, constructing a good explanation

requires the creation of new concepts and/or tools which would elucidate the phenomenon or problem circumscribed by the explanandum. For instance, explaining certain unexpected cognitive patterns which arise in the context of rapid learning tasks has required the development of specific probabilistic concepts which have been explored and refined by various connectionist and dynamicist models (cf. Guastello and Pincus 2009; McClelland et al. 2010). Similarly, there are certain neurobiological mechanisms that are arguably amenable to an explanation in terms of computational operations defined over functionally relevant units such as spike trains and/or firing rates (cf. Dayan and Abbott 2005; Ermentrout and Terman 2010).

The fact that explanation sometimes requires the construction of novel explanatory structures/concepts also vindicates the intuition that sometimes the request for explanation can remain unsatisfied. This can happen either because the way in which the explanandum is circumscribed does not enable the creation of an explanatory link, or because of the failure to provide an explanatory structure that elucidates the problem raised by the explanandum. Furthermore, within the broadly epistemic perspective in which I have situated the analysis of the notion of cognitive explanation, it is possible that certain things are not amenable to explanation at all. In other words, the view being proposed entails that there are things (aspects of reality) which are 'partially inscrutable' (Moravcsik 1998). This remark should not lead to any form of exaggerated skepticism, subjectivism, or relativism. For, as noted above, partial inscrutability of cognitive phenomena is compatible with the idea that it might be possible to develop conceptual and experimental tools for investigating certain aspects of cognition that at present are beyond our grasp. This idea could be strengthened by adding a historical dimension which shows that 'there are modes of understanding and patterns of explanation that at a previous point in time and conceptual context we would not have regarded as intelligible and yielding insight.' (ibid.: 191; cf. Hacking 1982, 2012; Williams 2002). Thus, the point of emphasising the partial inscrutability of cognition is not to give up the claim of objectivity, either for our general explanatory schemas or for the evaluation of what counts as insightful and yielding understanding. Rather, the underlying thought is that there is no way to determine in advance (of scientific research or human inquiry more generally) what can be known and understood about the nature and structure of cognitive phenomena.

There are two other important features of this pluralist model of cognitive explanation which help qualify further the sense in which cognitive explanations are essentially partial. Firstly, as mentioned above, a key intuition driving the arguments proposed so far is that in the case of a successful explanation, the explanans is typically taken to increase the intelligibility of the explanandum, thus making it less

problematic. Secondly, this way of representing the cognitive effects of offering a good (adequate) explanation implies that one should be able to connect the knowledge/information concerning the explanandum with that conveyed by the explanans. That is, I claim that explanation involves selecting, structuring, and organising the information or knowledge available about a given explanandum with the help of the conceptual structures made available by the explanans in a way that would yield further insight and understanding of the target problem/phenomenon.

Thus, at a very general level, the process of asking and giving explanations can be divided in the following four stages: (i) assume that certain things are true, (ii) construct some things as being problematic, (iii) take certain structures as having explanatory power, and (iv) apply these explanatory structures to elucidate the things which are deemed to be problematic and satisfy the request for explanation (cf. Moravcsik 1998, p. 161). At this point, a sceptic might argue that the picture being put forward does not address the issue of what determines the explanatory power of a particular conceptual structure. In response I contend that there is no unique factor which can establish whether a certain structure has an explanatory value or not. If the arguments against the monist commitments implicit in the existing accounts of cognitive explanation are on the right track, then a more fruitful way of addressing the sceptic's challenge is to critically analyse the various types of norms which govern the day-to-day practice of constructing good cognitive explanations.

From a pluralistic perspective explanation is viewed as a way of establishing connections between our different ways of conceptualising the world, and, as such, is essentially an epistemic activity. Otherwise put, explanations are primarily concerned with establishing relations between the concepts we develop in order to think about the world. However, at this point, one might insist that this notion of explanation makes sense only if one has already established a hierarchy of concepts. Whilst I think that this intuition is correct, it does not necessarily imply that there is a single way of determining, within a specific research programme or theory, a hierarchy of explanatory concepts. Equally, it does not imply that there are an indeterminate number of ways of specifying such conceptual hierarchies. Rather, I am defending a moderate view which acknowledges that there might be several ways in which to order the explanatory structures made available within a particular theoretical framework (or system of knowledge). Thus, in what follows, I will characterise explanations in terms of relations between concepts or ways of conceptualising reality, insisting that this sort of talk is meant to emphasise the epistemic (rather than the subjective) character of explanation.

However, this general account of what is involved in the process of constructing good cognitive explanations should not be confounded

with yet another form of explanatory monism. The idea that successful explanations create links between things that we find problematic (against the background of a given body of accepted scientific knowledge) and (new) structures which increase our intellectual grasp of them, merely sets the stage for a more detailed analysis of the various explanatory schemas proposed by different groups of practicing scientists. It is the careful scrutiny of the epistemic activities of constructing mechanistic, computational, or connectionist models/theories of cognitive phenomena which reveals the norms that play an essential role in establishing the explanatory power of the models proposed in each of these modelling paradigms.

For instance, among the mechanistic models of long term potentiation (Craver 2007b; Ermentrout and Terman 2010), the ones that are considered to have genuine explanatory power are those that are appropriately constrained by a series of ontological principles concerning the properties, activities, and organisation of the elements involved in the transmission of electrical neural signals. These principles are in turn reinforced by the available techniques of localisation, control, and manipulation of elements such as activating and inactivating molecules. In addition, in the construction of potentially explanatory models of LTP, one relies on bodies of accepted knowledge concerning processes such as vesicular packaging, the influence of inhibitors and agonists on the pre- and post-synaptic cells, and so on. Similarly, in the case of mechanistic models of ganglion cells' response profiles in the early visual pathway (Einevoll and Heggelund 2000; Einevoll and Plesser 2005; Kaplan and Craver 2011), we have seen that the construction of good mechanistic models is guided by a series of ontic and epistemic principles that also function as local norms.

However, against the strong mechanistic contention according to which these norms should apply to the evaluation of any proposed model/theory of a cognitive phenomenon, I have argued that cognitive models developed at different levels of analysis and/or abstraction are constrained by different sets of local norms and principles. For example, in the case of the computational models of early visual processing and object recognition, analysed in chapter 4, strong equivalence criteria such as response time profiles and other complexity measures, as well as the availability of appropriate semantic interpretations of the component computational structures, play an important role in establishing the adequacy and explanatory value of particular computational models of cognitive capacities (Shadmehr and Wise 2005; Sinha and Balas 2008). A similar conclusion was derived from the analysis of connectionist approaches to cognition, where the availability of semantic interpretations of the tasks performed by connectionist networks and considerations pertaining to the complexity and structure of these networks have been shown to play a crucial

role in determining the adequacy of particular types of networks for modelling certain aspects of cognitive processing such as the gradual character of language acquisition or language change (Thomas and McClelland 2008).

In brief, in the case of each explanatory framework it was possible to identify a set of local principles or norms which qualifies what counts as an explanatory structure with respect to the phenomenon being investigated (e.g., the molecular mechanisms involved in LTP, the ratio-template features of object recognition, the learning function characteristic of the process of linguistic inflection). However, the discussion of the examples from the mechanistic and computational modelling literature reinforces a further important point, namely that the explanatory link established between particular cognitive phenomena/patterns and specific mechanistic, computational or connectionist structures is often not graspable all at once or as a whole. This further suggests that the evaluation of the explanatory value of a particular model/theory requires taking into consideration a host of factors which cannot be reduced to a single category, as suggested by the philosophical models of cognitive explanation analysed in chapters 3 to 6.

In light of these considerations, I claim that the proposed conception also satisfies the requirement of accounting for the normative dimension of cognitive explanation. For on this account, there is a wide variety of principles and norms (ontic and epistemic) which guide the specific modelling and theorising activities of practising cognitive scientists. In particular, I have argued that mental content ascriptions play a crucial role in connecting abstract (computational or mathematical) explanatory structures to the explananda of cognitive theories. Given these qualifications concerning the normative dimension of cognitive explanation, the pluralist conception being put forward should be clearly distinguished from any reckless relativism.

According to a strong relativist, there is no point in evaluating the explanatory value of various scientific hypotheses and/or theories since these will always be relative to a particular set of aims and goals which will in turn vary from one scientific community to another. Thus, strong relativism undermines the very idea of using explanatory power as a criterion for comparing alternative theories and/or models of specific cognitive phenomena. In contrast, the form of explanatory pluralism I defend is incompatible with the renouncement of critical judgment concerning the explanatory value of alternative scientific theories. I have shown that, both from a principled and practice-based perspective, it is possible to identify a series of norms which guide the activities of constructing and evaluating adequate explanatory models/theories of cognitive phenomena.

What is being put forward, therefore, is not merely a descriptive account of the different explanatory practices currently encountered in

the branches of cognitive science, but a normative view of how these various practices should be understood and evaluated. I propose two additional motivations in support of this contention. The first relies on a general consideration concerning the organisation and structure of cognitive scientific research. Pluralism can be considered both a 'fact', which merely reflects the interdisciplinary and multilevel organisation of the disciplines of cognitive science, and a governance principle of the epistemic activities pursued by cognitive scientists. According to the second interpretation of pluralism, one should promote the cultivation of multiple systems of scientific knowledge, i.e., theories, research programmes (Mitchell 2009; Chang 2012). Along these lines, explanatory pluralism claims that scientific practice should encourage and support the development of a plurality of explanatory strategies in the investigation of cognitive phenomena. That is not to say that individual researchers or even research communities should pursue as many explanatory frameworks as possible, but rather that the field as a whole should promote the concurrent development of such schemas.

Second, the adoption of explanatory pluralism can be said to have a number of practical advantages. For instance, by allowing multiple explanatory strategies to be pursued simultaneously, one is better insured against unpredictability, and better poised to compensate for the limitations of each explanatory strategy when applied to the study of different aspects of cognition. In addition, explanatory pluralism supports the local integration of different explanatory hypotheses pertaining to specific cognitive phenomena, the co-optation of different elements (concepts) across explanatory schemas, and even the productive competition between explanatory tools and strategies. At a more general level, pluralism aims to direct the focus of philosophical analysis towards exploring the strengths and limitations of individual explanatory frameworks rather than constructing arguments for the ascendancy of one framework over another. This, I claim, will avoid much unproductive polarisation in the field and will perhaps even promote a more fruitful dialogue between different paradigms developed to study cognitive phenomena.

More generally, the conception of cognitive explanation being put forward can be characterised along broadly coherentist lines. This is because the view highlights the fact that, in constructing scientific explanations of cognitive phenomena, one starts from what is known without requiring that the starting points be absolutely secure. Otherwise put, the proposed pluralist view mirrors the practice of constructing and justifying the value of explanatory models/theories of cognition without insisting that such epistemic activities should end in an unshakable foundation. The account recognises that one builds explanatory models/theories of particular cognitive capacities starting from the achievements of some actual past group of cognitive

researchers. However, this coherentist insight is compatible with the fact that the same system of scientific knowledge can give rise to many alternative lines of inquiry and explanatory frameworks which need not always be judged in relation to each other. This type of pluralism also reinforces the idea that has surfaced before that different explanatory frameworks provide only *partial* accounts of the cognitive phenomena being investigated.

Throughout the thesis I have distinguished two senses in which proposed explanations of cognitive phenomena can be said to be partial. Using a spatial metaphor, one could say that explanations are 'vertically' partial if they do not provide the most detailed accounts possible of the fundamental feature revealed by the explanans (e.g., mechanistic sketches, simplified connectionist models), but they could be completed in light of further hypotheses and evidence, or by integrating hypotheses proposed at different levels of analysis or abstraction. More importantly from a pluralist perspective, explanations can also be partial in a 'horizontal' sense because they target certain problems which will be clarified in light of some specific system of knowledge. Granting that there is no ultimate and definitive way of establishing the unique, best system of knowledge, one must accept that there are multiple potentially explanatory accounts that target the same type of cognitive phenomena. Sometimes understanding a given class of phenomena is achieved precisely by applying concomitantly different models which generate multiple insights about various aspects of the phenomena being investigated. In light of these considerations, I would like to discuss next two further consequences of adopting pluralism as both a descriptively adequate and normative position with respect to the problem of cognitive explanation.

7.3.2 *Explanatory pluralism, unification, and realism*

In introducing *the problem of cognitive explanation* in chapter 2, I have pointed out that there are two traditional challenges that have been raised in relation to it: (i) the idea of a unified theory of cognition and (ii) the realisation problem. In what follows I aim to show how the thesis of explanatory pluralism fares with respect to these two challenges. I start by considering the issue of a unified theory of cognition, and then, in connection with the realisation challenge, discuss briefly the realist commitments of the account of cognitive explanation articulated above. The unification issue arises in the philosophy of psychology and cognitive science under a variety of guises. As we have seen, almost all of the theoretical frameworks claim that they are in principle able to unify/integrate hypotheses proposed within various sub-branches of cognitive science. This, it has been argued, reflects the prevalent monist assumption implicit in much cognitive scientific theorising and philosophising, according to which there is

only one 'right' framework that can deliver the true or correct theory of cognition. Even more sophisticated conceptions of the integrationist ideal of cognitive science (cf. Craver 2007b) fall into the same monist temptation and argue that there is a unique framework (e.g., mechanism) which guarantees the desired integration of the various hypotheses developed in different disciplines of cognitive science.

Two very general intuitions seem to drive and motivate the search for a unified theory of cognition. The first is a form of 'methodological optimism', according to which once you have found a particular set of concepts and/or experimental tools that help advance your investigation in a particular empirical domain, you have good reasons to expect that the strategy would work equally well (i.e., will deliver interesting results) when applied to another aspect/part of reality. This intuition can motivate in part the search for local unifications of different hypotheses developed at distinct levels of analysis but which target the same empirical (cognitive) phenomenon. Moreover, it is compatible with the proposed view, insofar as one of the advantages of developing multiple concurrent explanatory strategies is that, when it comes to it, one can co-opt different elements/tools across explanatory frameworks in order to obtain a new, more unified explanation of a particular target phenomenon.

Therefore, explanatory pluralism does not deny the possibility of local unifications, being consistent with the idea that achieving more unified theories is one of the multiple aims of scientific inquiry in the cognitive domain. However, it does insist on the separability (logical independence) of the two epistemic virtues: the unificatory power and the explanatory power of cognitive theories. Otherwise put, the version of explanatory pluralism defended above opposes the adoption of unification (or unificatory power of a theory) as a criterion for determining the explanatory value of a particular cognitive theory/model. In this way, the account avoids the main problems facing *explanatory unification* (Friedman 1974; Kitcher 1981, 1989): spurious and exclusionary unifications. Spurious unifications trivially reduce the number of explanatory frameworks without any real epistemic gain, whereas exclusionary unifications amount to thinking that the success of a particular scientific theory (e.g., string theory, molecular neurobiology) would leave no worthwhile scientific questions unanswered or would rule out the usefulness of other scientific theories. In contrast, the account put forward entails that both (local) unification as well as the compartmentalisation of scientific inquiry can and should be pursued for their explanatory value in the domain of cognitive science.

The second intuition supporting the ideal of a unified science of cognition is that, despite its apparent complexity, the world is ultimately unified, so that our explanations should in the long run capture the real unified structure of reality. In response, a number of

authors (e.g., Anderson 1972; Hacking 1983; Dupré 1993; Cartwright 1999; Laughlin and Pines 2000) have argued that 'the world is diverse and disorderly and that, correspondingly, the sciences are many and particular' (cf. Kitcher 1999, p. 338). A less radical reading of the main lesson of the various 'disunity' arguments is that it is an open question how much order or unity there is in nature and how we ought to proceed if what we are after is a completely unified account of nature. Explanatory pluralism should not be taken to decide the question of whether the world (or even a smaller part of it, viz. the cognitive domain) is ultimately unified or not. Rather, the pluralist thesis is compatible with the idea that we should seek to unify our scientific theories and hypotheses only so far as the structure of the world admits. Thus, by pursuing explanatory pluralism one does not deny that the world participates in determining what counts as explanatory. But given that we cannot know in advance how unified or complex the world really is, the best bet seems to be to cultivate multiple concurrent strategies for explaining and understanding different aspects of it.

There is yet another way to interpret the unificationist intuition. Rather than starting from an ontological assumption about the ultimate structure of the world, one might derive the purported unity of the world from a normative picture of scientific knowledge. For instance, assuming that scientific knowledge can be organised as a deductive system in which explanations of both general patterns and particular phenomena can be derived from first principles seems to support the idea that the world itself is thus organised and unified. Although the present investigation does not entail any general picture of scientific knowledge, the conception of explanation formulated in the previous section provides good reasons for doubting that it is possible to derive straightforwardly any unified and orderly picture of the world from our varied epistemic activities. Moreover, the burden of proof will fall on those who wish to maintain that the unity of the world can be derived from a general theory of scientific knowledge, given that the latter type of account does not seem to be forthcoming.

In summary, explanatory pluralism is neutral with respect to the prospects of developing a unified theory of cognition. The proposed analysis of the structure of cognitive explanation is perfectly compatible with the fact that (local) theoretical unifications can be achieved in a variety of ways, sometimes by reducing explanatory schemas and at other times by combining them in complex explanatory strategies which elucidate different aspects/features of particular cognitive phenomena. However, explanatory pluralism cannot and need not establish in advance how much unification will be achieved in the study of cognitive phenomena. This line of argument reinforces the idea that within a pluralistic framework, unification can and should be conceived as a separate aim of scientific inquiry, i.e., one which is

relatively independent from the explanatory power of scientific theories/models.

I turn next to the question of scientific realism. Within philosophy of cognitive science, this issue standardly takes the form of the so-called *realisation problem*, according to which cognitive explanations must refer to real biological mechanisms that support or maintain the cognitive phenomena being investigated. I have argued that there are three major concerns with the way in which the realisation problem has been framed and discussed in the literature. First, the realisation problem leaves open the question at which level one is supposed to find the 'real' physical mechanisms responsible for the cognitive phenomena being investigated. The possible candidates include the cellular level, the molecular, the biochemical level, and even deeper physical levels. Although one can postulate metaphysically (*a priori*) that there must be an ultimate level comprising the real mechanisms that underlie cognitive phenomena, from the perspective of our epistemic practices this solution will not be helpful at all in determining whether or not our current theories meet the realisation challenge. For it will still be indeterminate at what level one must pursue the construction of realistic cognitive explanations. Second, the realisation problem is actually too poorly-understood to count as the ultimate criterion for evaluating the merits of alternative theories of cognition (cf. Marcus 2001). And third, the realisation problem, at least in most of its current formulations, seems to imply that there should be a unique method of establishing when there is a correspondence between the theoretical posits of a particular cognitive theory/model and certain real physical entities or processes. The burden of proof seems to fall on those who wish to maintain that such a method is attainable given that it is not clear how such a method is to be derived.

However, I do not take this general criticism of the realisation problem to show that ontological principles (or realist commitments towards the existence of certain entities and/or processes) do not play any role in constructing (good) explanatory models of cognitive phenomena. On the contrary, as illustrated in the previous chapters of the thesis, ontological principles do play an important role in developing adequate explanatory frameworks for the study of cognitive phenomena. Moreover, I maintain that a careful analysis of these principles shows that there is a close and intricate relationship between the explanatory value of cognitive theories and their empirical adequacy.

The moderate form of scientific realism that underlies the proposed pluralist account of cognitive explanation amounts to claiming that ontological considerations should not be formulated *in abstracto* but must be founded on the experimental and theorising activities of the relevant scientific communities. As such they can be shown to guide/constrain the construction and refinement of explanatory mod-

els of particular cognitive capacities. This remark concludes the investigation of the consequences of adopting pluralism as the general framework for analysing the notion of cognitive explanation. In what follows, I provide some brief motivation for thinking that explanatory pluralism might represent an appropriate frame of reference for other areas of scientific investigation as well.

7.4 EXPLANATORY PLURALISM BEYOND COGNITIVE SCIENCE

The general account of cognitive explanation put forward in this thesis claims that there is no logical structure or abstract category that by itself is able to define what counts as explanatory. A careful look at cognitive scientific practices shows that the elements of the explanatory complexes used to elucidate particular aspects of cognitive phenomena change, which in turn supports the idea that we do not have an adequate way of characterising for all times and contexts the elements of something explanatory. However, this does not imply any radical skepticism towards the philosophical project of analysing the structure of cognitive explanation.

Focusing on the notion of explanatory structure, the analyses developed in chapters 3 to 6 have remained as neutral as possible with respect to any specific metaphysical hypotheses concerning the nature and structure of the mind. The model provided in this thesis highlights two abstract features of cognitive explanations, namely that: (i) explanatory structures help to elucidate certain aspects of cognitive phenomena that are deemed to be problematic, and (ii) explanations create links between different ways of conceptualising certain parts of the world, thus yielding new insights and understanding of the phenomena being investigated. I have supported this general proposal in two different ways. First, I have attempted to prove its *cogency* by abstract considerations and arguments. Second, I have demonstrated the *applicability* of the account by showing that it can be used to clarify certain aspects of the modelling and theorising practices deployed in different disciplines of cognitive science. Hence, I have defended explanatory pluralism as a position that is both theoretically sound and adequate with regard to the field of cognitive science.[4]

This characterisation of the process of constructing good explanations does not seem to be restricted solely to the cognitive domain. In fact, given its abstract character, I claim that the explanatory pluralist thesis might be productively applied in order to evaluate the explanatory strategies utilised in other areas of scientific investigation as well.

However, one might object that whilst explanatory pluralism reflects in an appropriate way the diversity of schemas used in the do-

4 An even better sense of the general extent of its applicability would require further investigation of the explanatory frameworks utilised in other disciplines of cognitive science.

main of cognitive science, which is a relatively new and immature area of scientific inquiry, it would fail to be adequate with respect to a more mature scientific field such as biology, chemistry or physics. Appeals to the maturity of a science over another are standardly used to emphasise something about the stability of the results obtained in a particular field of inquiry as opposed to some other field (cf. Psillos 1999; Chang 2012). As such, they do not necessarily conflict with the general commitments of explanatory pluralism. For physics might be said to be mature in virtue of the stability of its results and still cultivate a wide range of explanatory frameworks which target phenomena at different levels of analysis or resolution (e.g., in thermodynamics, statistical mechanics, electromagnetism, quantum mechanics, etc.).

Nor should explanatory pluralism be taken to downplay the significant dissimilarities between different areas of scientific investigation. It is perfectly possible that the diversity of explanatory methods and tools used in cognitive science is, in some respects, due to the immaturity of the field. However, the present account does not saddle any particular field of scientific inquiry with a specific number of explanatory categories. In fact, such a prescription would go against the dynamic character of the pluralist account advocated in this thesis. Furthermore, this version of explanatory pluralism is consistent with the idea that, despite a range of relevant variations, different fields of scientific inquiry do sometimes use very similar conceptual tools to investigate and explain very different or disparate aspects of the world.

At the most general level, this view receives *prima facie* support from the observation of *de facto* plurality of models, theoretical approaches, experimental techniques, and explanations utilised in different areas of scientific inquiry. In fact, pluralism has been defended within general philosophy of science from various perspectives by a number of authors (e.g., Hacking 1983, 2012; Dupré 1993; Cartwright 1999; Mitchell 2003; Chang 2012). Beyond these motivational remarks, further support for the idea that explanatory pluralism applies to other areas of scientific inquiry besides cognitive science would be gained through a detailed and rigorous analysis of the relevant theoretical and experimental practices. However, this sort of analysis falls outside the scope of this thesis which has focused on defending explanatory pluralism with respect to the domain of cognitive science.

BIBLIOGRAPHY

Abrahamsen, A. and W. P. Bechtel (2006). "Phenomena and Mechanisms: Putting the Symbolic, Connectionist, and Dynamical Systems Debate in Broader Perspective". In: *Contemporary Debates in Cognitive Science*. Ed. by R. Stainton. Basil Blackwell (cit. on pp. 56, 59).

Achinstein, P. (1983). *The Nature of Explanation*. Oxford University Press (cit. on p. 35).

Anderson, P. W. (1972). "More is Different". *Science* CLXXVII, pp. 393–396 (cit. on p. 221).

Batterman, R. W. (2000). "A 'Modern' (= Victorian ?) Attitude Towards Scientific Understanding". *The Monist* 83.2, pp. 228–257 (cit. on p. 139).

— (2002). *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*. Oxford University Press (cit. on pp. 21, 22, 139).

— (2010). "On the Explanatory Role of Mathematics in Empirical Science". *The British Journal for the Philosophy of Science* 61, pp. 1–25 (cit. on p. 22).

Bechtel, W. (2008). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. Oxford: Routledge (cit. on pp. 12, 50, 51, 58, 59, 68, 194, 197).

— (2011). "Mechanism and Biological Explanation". *Philosophy of Science* 78, pp. 533–557 (cit. on pp. 50, 59, 66, 68, 197).

Bechtel, W. and A. Abrahamsen (1991). *Connectionism and the Mind: An Introduction to Parallel Processing in Networks*. Basil Blackwell (cit. on pp. 152, 155, 156, 161, 164, 208).

— (2010). "Dynamic Mechanistic Explanation: Computational Modeling of Circadian Rhythms as an Exemplar for Cognitive Science". *Studies in History and Philosophy of Science Part A* 41.3, pp. 321–333 (cit. on pp. 56, 59).

— (2013). "Thinking Dynamically About Biological Mechanisms: Networks of Coupled Oscillators". *Foundations of Science* 18.4, pp. 707–723 (cit. on pp. 56, 59).

Bechtel, W. and R.C. Richardson (1993/2010). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. 2nd. MIT Press/Bradford Books (cit. on pp. 48, 50, 58, 68, 127, 172, 182, 194, 204, 208).

Bechtel, W. and C. D. Wright (2007). "Mechanisms and Psychological Explanation". In: *Philosophy of Psychology and Cognitive Science*. Ed. by P. Thagard. Elsevier (cit. on pp. 50, 59).

Bechtel, W. and C. D. Wright (2009). "What is Psychological Explanation?" In: *Routledge Companion to the Philosophy of Psychology*. Ed. by P. Calvo and J. Symons. Routledge (cit. on p. 50).

Bell, A. J. and T. Sejnowski (1997). "The 'Independent Components' of Natural Scenes are Edge Filters". *Vision Research* 37.23, pp. 3327–3338 (cit. on p. 93).

Berwick, R. C. et al. (2011). "Poverty of the Stimulus Revisited". *Cognitive Science* 35.7, pp. 1207–1242 (cit. on p. 149).

Block, N. (1986). "Advertisment for a Semantics for Psychology". In: *Midwest Studies in Philosophy: Studies in the Philosophy of Mind*. Ed. by P. French et al. University of Minnesota Press (cit. on pp. 40, 84).

— (1997). "Anti-reductionism Slaps Back". *Philosophical Perspectives* 11, pp. 107–32 (cit. on pp. 131, 199).

Bogen, J. (2008). "The Hodgkin Huxley Equations and the Concrete Model: Comments on Craver, Schaffner, and Weber". *Philosophy of Science* 75.5, pp. 1034–1046 (cit. on pp. 195, 196).

Bokulich, A. (2011). "How Scientific Models Can Explain". *Synthese* 180, pp. 33–45 (cit. on pp. 32, 33).

Broadbent, D. E. (1954). "The Role of Auditory Localization in Attention and Memory Span". *Journal of Experimental Psychology* 47, pp. 191–6 (cit. on pp. 38, 39).

Bromberger, S. (1992). *On What We Know We Don't Know: Explanation, Theory, Linguistics and How Questions Shape Them*. University of Chicago Press (cit. on p. 35).

Burge, T. (1979). "Individualism and the Mental". *Midwest Studies in Philosophy* 4.1, pp. 73–122 (cit. on pp. 88, 92, 97, 201).

— (1986). "Individualism and Psychology". *Philosophical Review* 95, pp. 3–45 (cit. on pp. 82, 88, 92, 97, 201).

— (2010). *Origins of Objectivity*. Clarendon Press: Oxford (cit. on pp. 88, 97).

Cartwright, N. (1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge University Press (cit. on pp. 15, 221, 224).

Chalmers, D. (1996). *The Conscious Mind*. Oxford University Press (cit. on p. 84).

— (2002). "The Components of Content". In: *Philosophy of Mind: Classical and Contemporary Readings*. Oxford University Press (cit. on p. 84).

Chang, H. (2009). "Ontological Principles and the Intelligibility of Epistemic Activities". In: *Scientific Understanding: Philosophical Perspectives*. Ed. by H. de Regt, S. Leonelli, and K. Eigner. University of Pittsburgh Press (cit. on p. 9).

— (2012). *Is Water H2O?: Evidence, Realism, and Pluralism*. Springer (cit. on pp. 9, 15, 44, 218, 224).

Chemero, A. and M. Silberstein (2008). "After the Philosophy of Mind: Replacing Scholasticism with Science". *Philosophy of Science* 75.1, pp. 1–27 (cit. on pp. 15, 61).

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton (cit. on pp. 38, 39).

— (1959). "A Review of B. F. Skinner's Verbal Behavior". *Language* 35, pp. 26–58 (cit. on p. 38).

— (1975). "Cartesian Linguistics: Acquisition and Use of Language". In: *Innate Ideas*. Ed. by S. Stich. University of California Press (cit. on p. 149).

— (1995). "Language and Nature". *Mind* 104.413, pp. 1–61 (cit. on pp. 96, 102, 200, 201, 203).

Christiansen, M. H. and N. Chater (1994). "Generalisation and Connectionist Language Learning". *Mind and Language* 9.3, pp. 273–287 (cit. on p. 162).

— (2002). *Connectionist Psycholinguistics*. Ablex Publishing Corp. (cit. on pp. 149, 165, 178).

— (2008). "Language as Shaped by the Brain". *Behavioral and Brain Sciences* 31, pp. 489–558 (cit. on pp. 162, 165, 178).

Church, A. (1936). "An Unsolvable Problem in Elementary Number Theory". *The American Journal of Mathematics* 58, pp. 345–363 (cit. on p. 112).

Churchland, P. S. and T. Sejnowski (1992). *The Computational Brain*. MIT Press (cit. on p. 117).

Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again*. Cambridge University Press (cit. on p. 166).

Cohen, E. and P. Sterling (1991). "Microcircuitry Related to the Receptive Field Center of the On-Beta Ganglion-Cell". *Journal of Neurophysiology* 65.2, pp. 352–59 (cit. on p. 62).

Copeland, B. J. (1996). "What is Computation?" *Synthese* 108, pp. 224–359 (cit. on pp. 129, 130).

— (2002). "Hypercomputation". *Minds and Machines* 12.4, pp. 461–502 (cit. on p. 115).

Copeland, B. J. and D. Proudfoot (1996). "On Alan Turing's Anticipation of Connectionism". *Synthese* 108, pp. 361–377 (cit. on p. 208).

Copeland, B. J. and O. Shagrir (2011). "Do Accelerating Turing Machines Compute the Uncomputable?" *Minds and Machines* 21.2, pp. 221–239 (cit. on p. 115).

Crane, T. (2003). *The Mechanical Mind: A Philosophical Introduction to Minds, Machines, and Mental Representation*. Routledge (cit. on p. 18).

Craver, C. F. (2006a). "Physical Law and Mechanistic Explanation in the Hodgkin and Huxley Model of the Action Potential". *Philosophy of Science* 75.5, pp. 1022–1033 (cit. on p. 195).

— (2006b). "Why Mechanistic Models Explain". *Synthese* 153, pp. 355–376 (cit. on pp. 53, 54, 58, 68).

Craver, C. F. (2007a). "Constitutive Explanatory Relevance". *Journal of Philosophical Research* 32, pp. 3–20 (cit. on p. 195).

— (2007b). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press (cit. on pp. 12, 20, 21, 30, 31, 34, 48–51, 54, 55, 58, 59, 61, 66, 68, 70, 71, 77, 127, 131, 193–197, 199, 204, 216, 220).

— (2012). "Scientific Explanation: The Ontic Conception". In: *Explanation in the Biological and Historical Sciences*. Ed. by A. Hutterman and M. Kaiser. Berlin: Springer (cit. on pp. 30, 34, 55, 58, 59, 66, 68, 196, 197).

Craver, C. F. and W. Bechtel (2007). "Top-Down Causation without Top-Down Causes". *Biology and Philosophy* 22, pp. 715–734 (cit. on p. 128).

Craver, C. F. and G. Piccinini (2011). "Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches". *Synthese* 183.3, pp. 283–311 (cit. on pp. 13, 49–51, 53–55, 68, 70, 71, 77, 109, 111, 121, 131, 134, 139, 195, 199, 205).

Cummins, R. C. (1983). *The Nature of Psychological Explanation*. MIT Press (cit. on pp. 11, 40–42, 48, 163, 187, 199).

— (1989). *Meaning and Mental Representation*. MIT Press (cit. on pp. 13, 76, 77).

— (2000). "'How Does it Work' versus 'What are the Laws?' : Two Conceptions of Psychological Explanation". In: *Explanation and Cognition*. Ed. by F. Keil and R. A. Wilson. MIT Press (cit. on pp. 20, 40, 42, 48).

— (2010). *The World in the Head*. Oxford University Press (cit. on pp. 11, 41, 42, 44, 77, 187).

Dale, R., E Dietrich, and A. Chemero (2009). "Explanatory Pluralism in Cognitive Science". *Cognitive Science* 33.2, pp. 739–742 (cit. on p. 15).

Daugherty, K. G. et al. (1993). "Why No Mere Mortal Has Ever Flown Out to Center Field but People Often Say They Do". *Proceedings of the Fifteen Annual Conference of the Cognitive Science Society*, pp. 383–388 (cit. on p. 175).

Davies, M. (1991). "Individualism and Perceptual Content". *Mind* 100.399, pp. 461–484 (cit. on p. 92).

Dawis S., R. Shapley E. Kaplan and D. Tranchina (1984). "The Receptive Field Organization of X-Cells in the Cat: Spatiotemporal Coupling and Asymmetry". *Vision Research* 24.6, pp. 549–64 (cit. on p. 63).

Dayan, P. and L. F. Abbott (2005). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press (cit. on pp. 52, 214).

Dretske, F. (1988). *Explaining Behavior*. MIT Press/Bradford Books (cit. on p. 40).

Dupré, J. (1993). *The Disorder of Things*. Harvard University Press (cit. on pp. 9, 15, 221, 224).

Dyer, M. G. (1991). "Connectionism versus Symbolism in High-Level Cognition". In: *Connectionism and the Philosophy of Mind*. Ed. by T. E. Horgan and J. L. Tienson. Kluwer (cit. on p. 165).

Edgington, D. (2011). "Causation First: Why Causation is Prior to Counterfactuals". In: *Understanding Counterfactuals, Understanding Causation: Issues in Philosophy and Psychology*. Ed. by McCormack T. Hoerl C. and S. R. Beck. Oxford University Press (cit. on p. 32).

Egan, F. (1992). "Individualism, Computation, and Perceptual Content". *Mind* 101, pp. 443–459 (cit. on pp. 13, 76, 89, 92, 96, 200, 201).

— (1995). "Computation and Content". *The Philosophical Review* 104, pp. 181–203 (cit. on pp. 76, 92).

— (1999). "In Defense of Narrow Mindedness". *Mind and Language* 14, pp. 177–194 (cit. on pp. 13, 89, 91, 97, 102, 103, 200–202).

— (2009). "Is There a Role for Representational Content in Scientific Psychology?" In: *Stich and His Critics*. Ed. by D. Murphy and M. A. Bishop. Wiley-Blackwell (cit. on p. 83).

— (2010). "Computational Models: A Modest Role for Content". *Studies in History and Philosophy of Science* 41, pp. 253–259 (cit. on pp. 13, 76, 89, 90, 92, 97, 101–104, 200–203).

— (2013). *How to Think about Mental Content* (cit. on pp. 89, 101, 102, 200).

Egan, F. and R. J. Matthews (2006). "Doing Cognitive Neuroscience: A Third Way". *Synthese* 153.3, pp. 377–391 (cit. on p. 42).

Einevoll, G. T. and P. Heggelund (2000). "Mathematical Models for the Spatial Receptive-Field Organization of Nonlagged X-Cells in Dorsal Lateral Geniculate Nucleus of Cat". *Visual Neuroscience* 17.6, pp. 871–85 (cit. on pp. 62, 216).

Einevoll, G. T. and H. E. Plesser (2005). "Responses of the Difference-of-Gaussians Models to Circular Drifting-Grating Patches". *Visual Neuroscience* 22.4, pp. 437–46 (cit. on pp. 63, 216).

Eliasmith, C. (2007). "How to Build a Brain: From Function to Implementation". *Synthese* 153.3, pp. 373–388 (cit. on p. 211).

— (2010). "How We Ought to Describe Computation in the Brain". *Studies in History and Philosophy of Science Part A* 41.3, pp. 313–320 (cit. on p. 211).

Eliasmith, C. and T. Stewart (2012). "Compositionality and Biologically Plausible Models". In: *Oxford Handbook of Compositionality*. Ed. by W. Hinzen, E. Machery, and M. Werning. Oxford University Press (cit. on pp. 160–162, 165).

Elman, J. et al. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. MIT Press (cit. on pp. 149, 150, 158, 172, 175, 177, 208).

Elman, J. L. (1990). "Finding Structure in Time". *Cognitive Science* 14.2, pp. 179–211 (cit. on pp. 155, 172, 177).
— (1991). "Distributed Representations, Simple Recurrent Networks, and Grammatical Structure". *Machine Learning* 7.2-3, pp. 195–225 (cit. on pp. 155, 177).
— (1993). "Learning and Development in Neural Networks: The Importance of Starting Small". *Cognition* 48.1, pp. 71–99 (cit. on pp. 155, 177).
Ermentrout, G.B. and D. H. Terman (2010). *Mathematical Foundations of Neuroscience*. Springer (cit. on pp. 196, 214, 216).
Faye, J. (2007). "The Pragmatic-Rhetorical Theory of Explanation". In: *Rethinking Explanation*. Ed. by J. Persson and P. Ylikoski. Springer (cit. on p. 21).
Field, H. (1978). "Mental Representation". *Erkenntnis* 13 (cit. on p. 40).
Fodor, J. (1968). *Psychological Explanation: An Introduction to the Philosophy of Psychology*. Random House (cit. on p. 40).
— (1974). "Special Sciences and the Disunity of Science as a Working Hypothesis". *Synthese* 28, pp. 77–115 (cit. on pp. 11, 20, 40, 42, 131, 195, 199).
— (1975). *The Language of Thought*. Harvard University Press (cit. on pp. 13, 76, 80, 82, 85).
— (1980). "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology". *Behavioral and Brain Sciences* 3 (cit. on pp. 12, 76, 86, 87, 89, 134, 200, 201).
— (1987). *Psychosemantics*. MIT Press/Bradford Books (cit. on pp. 40, 76, 84, 85).
— (1997). "Special Sciences: Still Autonomous After All These Years". *Philosophical Perspectives* 11, pp. 149–63 (cit. on pp. 131, 195, 199).
Fodor, J. and Z. Pylyshyn (1988). "Connectionism and Cognitive Architecture: Why Smolensky's solution Doesn't Work". *Cognition* 28.3-71 (cit. on pp. 76, 100, 115, 161–163, 166, 187).
Frank, S. L., W. F. G. Haselager, and I. van Rooij (2009). "Connectionist Semantic Systematicity". *Cognition* 110, pp. 358–379 (cit. on pp. 100, 159, 161, 167).
Friedman, M. (1974). "Explanation and Scientific Understanding". *Journal of Philosophy* 71.1, pp. 5–19 (cit. on pp. 10, 34, 220).
Gallistel, R. C. (1993). *The Organisation of Learning*. MIT Press (cit. on p. 200).
Gallistel, R. C. and A. King (2009). *Memory and the Computational Brain: Why Cognitive Science Will Transform Neuroscience*. Blackwell/Wiley (cit. on pp. 12, 13, 76, 179, 200).
Gazzaniga, M. S. (2000). *The New Cognitive Neurosciences*. Ed. by M. S. Gazzaniga. Second. MIT Press/Bradford Books (cit. on p. 47).
Glennan Stuart, S. (1996). "Mechanisms and the Nature of Causation". *Erkenntnis* 44.1, pp. 49–71 (cit. on pp. 30, 48, 128).

Glennan Stuart, S. (2002). "Rethinking Mechanistic Explanation". *Proceedings of the Philosophy of Science Association* 3, pp. 342–353 (cit. on pp. 28, 30, 34, 48, 50, 128).

— (2005). "Modelling Mechanisms". *Studies in History and Philosophy of Science Part C* 36.2, pp. 443–464 (cit. on p. 50).

— (2010). "Ephemeral Mechanisms and Historical Explanation". *Erkenntnis* 72.2, pp. 251–266 (cit. on p. 128).

Gopnik, A. and L. (eds.) Schulz (2007). *Causal Learning: Psychology, Philosophy, and Computation*. Oxford University Press (cit. on p. 32).

Griffiths, T. L., C. Kemp, and J. B. Tenenbaum (2008). "Bayesian Models of Cognition". In: *The Cambridge Handbook of Computational Psychology*. Ed. by Ron Sun. Cambridge University Press (cit. on p. 34).

Guastello S. J., Koopmans M. and D. Pincus (2009). *Chaos and Complexity in Psychology: The Theory of Nonlinear Dynamical Systems*. Cambridge University Press (cit. on pp. 34, 214).

Hacking, I. (1982). "Language, Truth, and Reason". In: Basil Blackwell. Chap. 2 (cit. on p. 214).

— (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press (cit. on pp. 221, 224).

— (2012). "'Language, Truth, and Reason' 30 years later". *Studies in History and Philosophy of Science* 43, pp. 599–609 (cit. on pp. 214, 224).

Hansson, B. (2006). "Why Explanations? Fundamental, and Less Fundamanetal Ways of Understanding the World". *Theoria* 72.1, pp. 23–59 (cit. on pp. 24, 34).

— (2007). "Explanations Are About Concepts and Concept Formation". In: *Rethinking Explanation*. Ed. by J. Persson and P. Ylikoski. Springer (cit. on p. 24).

Hare, M., J. Elman, and K. G. Daugherty (1995). "Default Generalization in Connectionist Networks". *Language and Cognitive Processes* 10, pp. 601–630 (cit. on pp. 175, 176).

Hebb, D. O. (1949). *The Organization of Behavior*. Wiley and Sons (cit. on pp. 153, 157).

Hempel, C. G. (1965). "Aspects of Scientific Explanation". In: *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. The Free Press, pp. 331–496 (cit. on pp. 10, 20, 21, 23, 24, 33, 193).

— (2001). *The Philosophy of Carl G. Hempel: Studies in Science, Explanation, and Rationality*. Ed. by J. H. Fetzer. Oxford University Press (cit. on p. 20).

Hempel, C. G. and P. Oppenheim (1948). "Studies in the Logic of Explanation". *Philosophy of Science* 15.2, pp. 135–175 (cit. on pp. 10, 20, 21, 193).

Hille, B. (2001). *Ion Channels of Excitable Membranes*. Sunderland: Sinauer (cit. on p. 196).

Hodgkin, A. L. and A. F. Huxley (1952). "A Quantitative Description of Membrane Current and its Application to Conduction and Excitation in Nerve". *Journal of Physiology* 116.500-544 (cit. on p. 195).

Hopfield, J. J. (1982). "Neural Networks and Physical Systems with Emergent Collective Computational Abilities". *Proceedings of the National Academy of Sciences of the USA* 79.8, pp. 2554–2558 (cit. on p. 156).

Hubel, D. H. and T. N. Wiesel (1962). "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex". *Journal of Physiology* 160, pp. 106–154 (cit. on p. 117).

Hummel, J. E. and K. J. Holyoak (1997). "Distributed Representations of Structure: A Theory of Analogical Access and Mapping". *Psychological Review* 104.3, pp. 427–466 (cit. on p. 160).

— (2003). "A Symbolic-Connectionist Theory of Relational Inference and Generalisation". *Psychological Review* 110.2, pp. 220–264 (cit. on p. 160).

Illari, P. (2013). "Mechanistic Explanation: Integrating the Ontic and Epistemic". *Erkenntnis* (cit. on pp. 60, 69, 198).

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press (cit. on pp. 161–163).

Jeffrey, R. (1969). "Statistical Explanation vs. Statistical Inference". In: *Essays in Honor of Carl G. Hempel*. Ed. by N. Rescher. Doredrecht: D. Reidel (cit. on pp. 10, 20, 23, 25, 33, 193).

Jones, M. et al. (1997). "Top-Down Learning of Low-Level Vision Tasks". *Current Biology* 7, pp. 991–994 (cit. on p. 95).

Kaplan, D. M. (2011). "Explanation and Description in Computational Neuroscience". *Synthese* 183, pp. 339–373 (cit. on pp. 48, 49, 68, 109, 139).

Kaplan, D.M and C. F. Craver (2011). "The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective". *Philosophy of Science* 78, pp. 601–627 (cit. on pp. 48–53, 55, 63, 64, 66, 68, 196, 197, 216).

Kelso, J. A. S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behvior*. MIT Press (cit. on p. 34).

Kim, J. (1994). "Explanatory Knowledge and Metaphysical Dependence". *Philosophical Issues* 5, pp. 51–69 (cit. on p. 21).

Kitcher, P. (1981). "Explanatory Unification". *Philosophy of Science* 48.4, pp. 507–531 (cit. on pp. 34, 220).

— (1989). "Explanatory Unification and the Causal Structure of the World". In: *Scientific Explanation*. Ed. by P. Kitcher and W. C. Salmon. University of Minnesota Press, pp. 410–505 (cit. on pp. 10, 21, 24, 30, 34, 220).

— (1999). "Unification as a Regulative Ideal". *Perspectives on Science* 7.3, pp. 337–348 (cit. on p. 221).

Kuffler, S. (1953). "Discharge Patterns and Functional Organization of Mammalian Retina". *Journal of Neurophysiology* 16.1, pp. 37–68 (cit. on p. 62).

Ladyman, J. and R. Brown (2009). "Physicalism, Supervenience, and the Fundamental Level". *Philosophical Quaterly* 59.234, pp. 20–38 (cit. on p. 56).

Laughlin, R. B. and D. Pines (2000). "The Theory of Everything". *Proceedings of the National Academy of Sciences of the USA* XCVII, pp. 28–31 (cit. on p. 221).

Laurence, S. and E. Margolis (2001). "The Poverty of the Stimulus Argument". *British Journal for the Philosophy of Science* 52.2, pp. 217–276 (cit. on p. 149).

Lewis, D. (1986). "Causal Explanation". In: vol. Philosophical Papers II. Oxford University Press (cit. on pp. 20, 28, 32, 34).

— (2001). *Counterfactuals*. 2nd. Blackwell Publishers (cit. on pp. 31, 32).

Machamer, P., L. Darden, and C. F. Craver (2000). "Thinking about Mechanisms". *Philosophy of Science* 67.1, pp. 1–25 (cit. on pp. 28, 30, 34, 48, 50, 128, 204).

MacWhinney, B. and J. Leinbach (1991). "Implementations are not Conceptualizations: Revising the Verb Learning Model". *Cognition* (cit. on p. 175).

Marcus, G. F. (1999). "Connectionism: With or Without Rules? Response to J. L. McClelland and D. C. Plaut". *Trends in Cognitive Science* 3.5, pp. 168–170 (cit. on p. 174).

— (2001). *The Algebraic Mind*. MIT Press (cit. on pp. 165, 173–176, 178, 179, 210, 222).

— (2013). "Tree Structure and the Representation of Sentences: A Reappraisal". In: *Birdsong, Speech, and Language: Exploring the Evolution of Mind and Brain*. Ed. by Johan J. Bolhuis and Martin Everart. MIT Press (cit. on p. 189).

Marcus, G. F. et al. (1992). "Overregularization in Language Acquisition". *Monographs of the Society for Research in Child Development* 57.228 (cit. on p. 173).

Marr, D. (1982). *Vision*. Freeman (cit. on pp. 13, 76, 107, 195, 199).

Matthews, R. J. (1994). "Three-Concept Monte: Explanation, Implementation, and Systematicity". *Synthese* 101, pp. 347–363 (cit. on p. 188).

— (1997). "Can Connectionists Explain Systematicity?" *Mind and Language* 12.2, pp. 154–177 (cit. on pp. 100, 188, 209).

— (2007). *The Measure of Mind: Propositional Attitudes and Their Attribution*. Oxford University Press (cit. on p. 201).

McClelland, J. L. et al. (2010). "Letting Structure Emerge: Connectionist and Dynamical Systems Approaches to Cognition". *Trends in Cognitive Science* 14, pp. 348–356 (cit. on pp. 34, 168, 172, 208, 210, 214).

McClelland, J. L. et al. (1995). "Why There Are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory". *Psychological Review* 102, pp. 419–457 (cit. on pp. 187, 211).

McCulloch, W. S. and W. H. Pitts (1943). "A Logical Calculus of the Ideas Immanent in Nervous Activity". *Bulletin of Mathematical Biophysics* 7, pp. 115–133 (cit. on pp. 116, 119, 120, 180).

McKenzie, K. (2011). "Arguing Against Fundamentality". *Studies in History and Philosophy of Science Part B* 42.4, pp. 244–255 (cit. on pp. 56, 57).

Mellor, D. H. (1976). "Probable Explanation". *Australasian Journal of Philosophy* 54.3, pp. 231–241 (cit. on pp. 10, 20, 23, 26, 27, 33, 193).

Mendola, J. (2008). *Anti-Externalism*. Oxford University Press (cit. on p. 84).

Milkowski, M. (2010). "Beyond Formal Structure: A Mechanistic Perspective on Computation and Implementation". *Journal of Cognitive Science* 12.4, pp. 359–379 (cit. on p. 109).

— (2013). *Explaining the Computational Mind*. MIT Press (cit. on p. 109).

Miller, G.A. (1956). "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information". *Psychological Review* 63, pp. 81–97 (cit. on pp. 38, 39).

Millikan, R. (1984). *Language, Thought and Other Biological Categories*. MIT Press/Bradford Books (cit. on p. 40).

Mills, J. W. (2008). "The Nature of the Extended Analog Computer". *Physica D: Nonlinear Phenomena* 237.9, pp. 1235–1256 (cit. on p. 117).

Minsky, M. and S. Papert (1972). *Perceptrons: An Introduction to Computational Geometry*. 2nd. MIT Press (cit. on p. 116).

Mitchell, S. D. (2003). *Biological Complexity and Integrative Pluralism*. Cambridge University Press (cit. on pp. 15, 44, 182, 212, 224).

— (2009). *Unsimple Truths: Science, Complexity, and Policy*. University of Chicago Press (cit. on p. 218).

— (2012). "Emergence: Logical, Functional, and Dynamical". *Synthese* 185, pp. 171–186 (cit. on pp. 44, 182).

Moravcsik, J. A. (1998). *Meaning, Creativity, and the Partial Inscrutability of the Human Mind*. CSLI Publications (cit. on pp. 214, 215).

Naundorf, B. F., F. Wolf, and M. Volgushev (2006). "Unique Features of Action Potential Initiation in Cortical Neurons". *Nature* 440, pp. 1060–1063 (cit. on p. 196).

Newell, A. (1980a). "Physical Symbol Systems". *Cognitive Science* 4, pp. 135–183 (cit. on pp. 76, 89).

— (1980b). "The Knowledge Level". *Americal Association for Artificial Intelligence* (cit. on p. 76).

Olshausen, B. A. and D. J. Field (1996). "Emergence of Simple Cell Receptive Field Properties by Learning a Sparse Code for Natural Images". *Nature* 381.6583, pp. 607–609 (cit. on p. 93).

Olshausen, B. A. and D. J. Field (2005). "How Close Are We to Understanding V1?" *Neural Computation* 17, pp. 1665–1699 (cit. on p. 93).

Papineau, D. (1987). *Reality and Representation*. Basil Blackwell (cit. on p. 40).

Peacocke, C. (1999). "Computation as Involving Content: A Response to Egan". *Mind and Language* 14.2, pp. 195–202 (cit. on pp. 81, 82).

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press (cit. on p. 31).

Phillips, C. (2013a). "On the Nature of Islands Constraints. I: Language Processing and Reductionist Accounts". In: *Experimental synatx and Island Effects*. Ed. by J. Sprouse and N. Hornstein. Cambridge University Press (cit. on p. 185).

— (2013b). "Some Arguments and Non-arguments for Reductionist Accounts of Syntactic Phenomena". *Language and Cognitive Processes* 28, pp. 156–187 (cit. on p. 185).

Phillips, C. and H Lasnik (2002). "Linguistics and Evidence: A Response to Edelman and Christiansen". *Trends in Cognitive Science*, p. 3 (cit. on p. 185).

Phillips, C. and S. Lewis (2013). "Derivational Order in Syntax: Evidence and Architectural Consequences". *Studies in Linguistics* 6, pp. 11–47 (cit. on p. 77).

Piccinini, G. (2004). "The First Computational Theory of Mind and Brain: A Close Look at McCullocj and Pitts's 'Logical Calculus of Ideas Immanent in Nervous Activity'". *Synthese* 141.2, pp. 175–215 (cit. on p. 122).

— (2007a). "Computational Modeling vs. Computational Explanation: Is Everything a Turing Machine, and Does It Matter to the Philosophy of Mind?" *Australasian Journal of Philosophy* 85.1, pp. 93–115 (cit. on pp. 111, 121, 122, 127, 128, 130, 205).

— (2007b). "Computationalism, the Church-Turing Thesis, and the Church-Turing Fallacy". *Synthese* 154.1, pp. 97–120 (cit. on pp. 13, 91, 109, 115, 117).

— (2008a). "Computation without Representation". *Philosophical Studies* 137.2, pp. 205–241 (cit. on pp. 13, 81, 97, 109, 111, 117, 121–125, 134, 139, 205).

— (2008b). "Some Neural Networks Compute, Others Don't". *Neural Networks* 21.2-3, pp. 311–321 (cit. on pp. 117, 122, 207).

— (2009). "Computationalism in the Philosophy of Mind". *Philosophy Compass* 4.3, pp. 515–532 (cit. on p. 111).

— (2010). "The Mind as Neural Software? Revisiting Functionalism, Computationalism, and Computational Functionalism". *Philosophy and Phenomenological Research* 81.2, pp. 269–311 (cit. on pp. 109, 111).

Piccinini, G. and S. Bahar (2013). "Neural Computationa and the Computational Theory of Cognition". *Cognitive Science* 37.3, pp. 453–

488 (cit. on pp. 13, 111, 113–117, 119, 120, 123, 131, 134, 180, 206, 207).

Pinker, S. (1995). *The Language Instinct*. Harper Perennial (cit. on p. 174).

Pinker, S. and A. Prince (1988). "On Language and Connectionism". *Cognition* 28.1-2, pp. 73–193 (cit. on p. 174).

Plate, T. (2003). *Holographic Reduced Representations*. Standford: CSLI Publications (cit. on p. 160).

Plunkett, K. and V. A. Marchman (1993a). "From Rote Learning to System Building: Acquiring Verb Morphology in Children and Connectionist Nets". *Cognition* 48, pp. 21–69 (cit. on p. 175).

— (1993b). "Learning from a Connectionist Model of the Acquisition of the English Past Tense". *Cognition* 61, pp. 299–308 (cit. on p. 175).

Port, R. and T. van Gelder (1995). *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT Press (cit. on p. 34).

Pour-El, M. B. (1974). "Abstract Computability and Its Relation to the General Purpose Analog Computer". *Transactions of the American Mathematical Society* 199, pp. 1–28 (cit. on p. 117).

Psillos, S. (1999). *Scientific Realism: How Science Tracks Truth*. Psychology Press (cit. on p. 224).

— (2002). *Causation and Explanation*. Acumen Publishing Ltd (cit. on p. 20).

— (2004). "A Glimpse of the Secret Connection: Harmonizing Mechanisms with Counterfactuals". *Perspectives on Science* 12.3, pp. 288–319 (cit. on p. 33).

— (2011). "The Idea of Mechanism". In: *Causality in the Sciences*. Ed. by Federica Russo Phyllis McKay Illari and Jon Williamson. Oxford University Press. Chap. 36, pp. 771–788 (cit. on pp. 50, 51, 197).

Putnam, H. (1975). *Mind, Language, and Reality*. Cambridge University Press (cit. on pp. 84, 129).

Pylyshyn, Z. W. (1984). *Computation and Cognition*. MIT Press (cit. on pp. 12, 13, 40, 76, 80, 85, 86, 89, 98, 99, 101, 106, 134, 200, 204).

Railton, P. (1978). "A Deductive-Nomological Model of Probabilistic Explanation". *Philosophy of Science* 45.2, pp. 206–226 (cit. on pp. 23, 28).

— (1981). "Probability, Explanation, and Information". *Synthese* 48, pp. 233–256 (cit. on pp. 23, 28, 29, 34).

— (1989). "Explanation and Metaphysical Controversy". In: *Scientific Explanation*. Ed. by P. Kitcher and W. C. Salmon. University of Minnesota Press (cit. on p. 61).

Ramsey, W. (1997). "Do Connectionist Representations Earn Their Explanatory Keep?" *Mind and Language* 12.1, pp. 34–66 (cit. on pp. 187, 209).

Ramsey, W. M. (2007). *Representation Reconsidered*. Cambridge University Press (cit. on pp. 163, 165, 187).

Rey, G. (1997). *Contemporary Philosophy of Mind: A Contentiously Classical Approach*. Blackwell Publishers (cit. on pp. 38, 130).

Rodieck, R. W. (1965). "Quantitative Analysis of Cat Retinal Ganglion Cell Response to Visual Stimuli". *Vision Research* 5.11, pp. 583–601 (cit. on pp. 62, 65, 196, 198).

Rubel, L. A. (1985). "The Brain as an Analog Computer". *Journal of Theoretical Neurobiology* 4, pp. 73–81 (cit. on p. 117).

— (1993). "The Extended Analog Computer". *Advances in Applied Mathematics* 14.1, pp. 39–50 (cit. on p. 117).

Rumelhart, D. E., J. L. McClelland, and the PDP Research Group (1986). *Parallel Distributed Processing*. Vol. 1: Foundations. MIT Press/Bradford Books (cit. on pp. 34, 149, 150, 152, 153, 156, 162, 164, 172, 174, 208).

— (1988). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Vol. 2: Psychological and Biological Models. MIT Press/Bradford Books (cit. on pp. 156, 180, 208).

Ryle, G. (1949/2002). *The Concept of Mind*. University of Chicago Press (cit. on p. 38).

Salmon, W. C. (1971). *Statistical Explanation and Statistical Relevance*. University of Pittsburgh Press (cit. on pp. 10, 20, 23, 25–28, 31, 33, 193).

— (1984a). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press (cit. on pp. 10, 20, 21, 28, 30, 31, 34, 58).

— (1984b). "Scientific Explanation: Three Basic Conceptions". *PSA: Proceedings of the Biennal Meeting of the Philosophy of Science Association* 2, pp. 293–305 (cit. on p. 34).

— (1989). "Four Decades of Scientific Explanation". In: *Scientific Explanation*. Ed. by P. Kitcher and W. C. Salmon. Vol. XIII. Minnesota Studies in the Philosophy of Science. University of Minnesota Press (cit. on pp. 10, 20, 24, 28, 34, 50, 51, 56, 193).

Schaffer, J. (2003). "Is There a Fundamental Level?" *Nous* 37, pp. 498–517 (cit. on p. 56).

Schreiner, C. E. and J. A. Winer (2007). "Auditory Cortex Mapmaking: Principles, Projections, and Plasticity". *Neuron* 56.2, pp. 356–365 (cit. on p. 117).

Searle, J. (1992). *The Rediscovery of the Mind*. MIT Press (cit. on p. 129).

— (2002). "Twenty-One Years in the Chinese Room". In: *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*. Ed. by J. M. Preston and M. A. Bishop. Oxford University Press (cit. on p. 129).

Shadmehr, R. and S.P. Wise (2005). *The Computational Neurobiology of Reaching and Pointing*. MIT Press/Bradford Books (cit. on pp. 47, 90, 212, 216).

Shagrir, O. (1997). "Two Dogmas of Computationalism". *Minds and Machines* 7.3, pp. 321–344 (cit. on p. 81).

Shagrir, O. (1999). "What Is Computer Science About?" *The Monist* 82.1, pp. 131–149 (cit. on p. 81).

— (2001). "Content, Computation, and Externalism". *Mind* 110.438, pp. 369–400 (cit. on pp. 76, 87, 91, 97, 159, 201, 202).

Silverberg, A. (2006). "Chomsky and Egan on Computational Theories of Vision". *Minds and Machines* 16.4, pp. 495–524 (cit. on pp. 92, 97).

Sinha, P. (2002). "Qualitative Representations For Recognition". In: *Lecture Notes in Computer Science*. Ed. by H. Bulthoff. Springer-Verlag (cit. on p. 94).

Sinha, P. and B. J. Balas (2008). "Computational Modeling of Visual Information Processing". In: *The Cambridge Handbook of Computational Psychology*. Vol. Ron Sun. Cambridge University Press (cit. on pp. 94, 95, 100, 216).

Smolensky, P. (1988a). "On the Proper Treatment of Connectionism". *Behavioral and Brain Sciences* 11.1, pp. 1–23 (cit. on pp. 161, 164).

— (1988b). "The Constituent Structure of Connectionist Mental States: A Reply to Fodor and Pylyshyn". *Southern Journal of Philosophy* 26.S1, pp. 137–161 (cit. on pp. 100, 161).

— (1990). "Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems". *Artificial Intelligence* 46, pp. 159–217 (cit. on p. 160).

Stainton, R. J. (2006). *Contemporary Debates in Cognitive Science*. Ed. by R. J. Stainton. Blackwell Publishers (cit. on p. 47).

Stalnaker, R. (1990). "Narrow Content". In: *Propositional Attitudes: The Role of Content in Logic, Language, and Mind*. Ed. by C. A. Anderson and J. Owens. Stanford: CSLI (cit. on p. 84).

— (1999). *Context and Content*. Oxford University Press (cit. on p. 84).

Stephan, A. (2006). "The Dual Role of 'Emergence' in the Philosophy of Mind and in Cognitive Science". *Synthese* 151, pp. 485–498 (cit. on p. 182).

Stepp, N., A. Chemero, and M. T. Turvey (2011). "Philosophy for the Rest of Cognitive Science". *Topics in Cognitive Science* 3.2, pp. 425–437 (cit. on pp. 15, 61, 63, 65).

Stich, S. (1983). *From Folk Psychology to Cognitive Science*. MIT Press Bradford Books (cit. on pp. 13, 83, 84, 201).

— (1991). "Narrow Content Meets Fat Syntax". In: *Meaning in Mind: Fodor and His Critics*. Ed. by B. Lower and G. Rey. Oxford, Blackwell (cit. on pp. 83, 84, 201).

Strevens, M. (2008). *Depth: An Account of Scientific Explanation*. Harvard University Press (cit. on pp. 30, 34, 55, 197).

Sun, Ron, ed. (2008). *The Cambridge Handbook of Computational Psychology*. Cambridge University Press (cit. on pp. 47, 77, 189).

Teuscher, C. (2002). *Turing's Connectionism: An Investigation of Neural Network Architectures*. Springer-Verlag (cit. on p. 208).

Thelen, E. and L. B. Smith (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. MIT Press (cit. on p. 34).

Thomas, M. S. C. and J. L. McClelland (2008). "Connectionist Models of Cognition". In: *The Cambridge Handbook of Computational Psychology*. Ed. by Ron Sun. Cambridge University Press (cit. on pp. 154, 211, 217).

Tolman, E.C., B.F. Ritchie, and D. Kalish (1946). "Studies in Spatial Learning, II: Place Learning versus Response Learning". *Journal of Experimental Psychology* 36, pp. 221–9 (cit. on p. 38).

Touretzky, D. S. and G. E. Hinton (1988). "A distributed Connectionist Production System". *Cognitive Science* 12.3, pp. 423–466 (cit. on pp. 165, 166).

Turing, A. M. (1937). "On Computable numbers, with an Application to the Entscheidungsproblem". *Proceedings of the London Mathematical Society* 2.42, pp. 230–265 (cit. on pp. 39, 112).

— (1948/2004). "Intelligent Machinery: A Report by A. M. Turing". In: *The Essential Turing*. Ed. by B. J. Copeland. Oxford University Press (cit. on pp. 115, 208).

van der Velde, F. and M. de Kamps (2006). "Neural Blackboard Architectures of Combinatorial Structures in Cognition". *Behavioral and Brain Sciences* 29.37-70 (cit. on p. 160).

van Fraassen, B. (1980). *The Scientific Image*. Oxford University Press (cit. on pp. 11, 21, 34, 35).

van Gelder, T. (1998). "The Dynamical Hypothesis in Cognitive Science". *Brain and Behavioral Sciences* 21, pp. 615–665 (cit. on p. 117).

Wagers, M. W. and C. Phillips (2007). "Relating Structure and Time in Linguistics and Psycholinguistics". In: *Oxford Handbook of Psycholinguistics*. Ed. by G. Gaskell. Oxford University Press (cit. on p. 185).

Werning, M. (2012). "Non-Symbolic Compositional Representation and its Neuronal Foundation: Towards an Emulative Semantics". In: *The Oxford Handbook of Compositionality*. Ed. by M. Werning and W. Hinzen. Oxford University Press, pp. 633–654 (cit. on pp. 161, 162, 165).

Williams, B. (2002). *Truth and Truthfulness: An Essay in Genealogy*. Princeton University Press (cit. on p. 214).

Wilson, M. (1985). "What Is This Thing Called 'Pain'? The Philosophy of Science Behind the Contemporary Debate". *Pacific Philosophical Quarterly* 66.227-67 (cit. on pp. 11, 41).

— (1992). "Frontier Law: Differential Equations and their Boundary Conditions". In: *P.S.A 1990*. Ed. by M. Forbes A. Fine and L. Wessels. Vol. II. E. Lansing: P. S. A. Press (cit. on p. 22).

— (2010). "Mixed-Level Explanation". *Philosophy of Science* 77.5, pp. 933–46 (cit. on pp. 22, 212).

Wilson, R. A. (1994). "Wide Computationalism". *Mind* 103, pp. 351–372 (cit. on pp. 82, 201).

Woodward, J. (2000). "Explanation and Invariance in the Special Sciences". *British Journal for the Philosophy of Science* 51.2, pp. 197–254 (cit. on pp. 12, 28, 31, 32, 48).

— (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press (cit. on pp. 20, 28, 31, 32, 34, 48, 49, 193).

— (2011). "Psychological Studies of Causal and Counterfactual Reasoning". In: *Understanding Counterfactuals, Understanding Causation: Issues in Philosophy and Psychology*. Ed. by McCormack T. Hoerl C. and S. R. Beck. Oxford University Press (cit. on p. 33).

Wright, C. D. (2012). "Mechanistic Explanation Without the Ontic Conception". *European Journal of Philosophy of Science* 2.3, pp. 375–394 (cit. on pp. 21, 58, 59, 197).

Yang, C. D. (2002). *Knowledge and Learning in Natural Language*. Oxford University Press (cit. on pp. 173, 176, 182, 184, 210).

— (2004). "Universal Grammar, Statistics or Both?" *Trends in Cognitive Science* 8.10, pp. 451–456 (cit. on p. 184).

Ylikoski, P. (2007). "The Idea of Contrastive Explanation". In: *Rethinking Explanation*. Ed. by J. Persson and P. Ylikoski. Springer (cit. on pp. 21, 32, 33, 35).

— (2013). "Causal and Constitutive Explanation Compared". *Erkenntnis* 78.2, pp. 277–297 (cit. on p. 195).

Ylikoski, P. and J. Kuorikoski (2010). "Dissecting Explanatory Power". *Philosophical Studies* 148, pp. 201–219 (cit. on pp. 32, 35).