Supervisor
Prof. Dr. Erik Weber
Co-Supervisor
Dr. Jeroen Van Bouwel

Reading Committee
Prof. Dr. Stuart Glennan, Butler University
Dr. Phyllis Illari, University College London
Dr. Bert Leuridan, Ghent University
Dr. Huib Looren de Jong, VU University Amsterdam
Prof. Dr. Joke Meheus, Ghent University

Dean
Prof. Dr. Marc Boone
Rector
Prof. Dr. Paul Van Cauwenberge

Nederlandse vertaling:
Verklaring in de cognitiewetenschappen en de biologie. Mechanismen, wetten en hun verklarende deugden.

Cover image: State-of-the-art map of the pigeon beak. The yellow structures were imaged using micro-computed tomography, which is used to capture dense structures like bone. The purple structures were produced with magnetic resonance imaging, showing the external soft tissues. Instead of magnetic neurons, the MRI image reveals iron-rich macrophages. *Image from UCL Centre for Advanced Biomedical Imaging, London, UK*.

ISBN: 1248852155445544

UNIVERSITEIT GENT

Faculteit Letteren & Wijsbegeerte

Raoul Gervais

Explanation in the cognitive sciences and biology

Mechanisms, laws, and their explanatory virtues

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de wijsbegeerte

2013

To the memory of my grandfather

Ernest Napoléon Raoul Gervais (1911-1993)

And therefore will I leave off metaphysical
Discussion, which is neither here nor there:
If I agree that what is, is; then this I call
Being quite perspicuous and extremely fair.

– Lord Byron, *Don Juan* (XI st. V)

Foreword

The present dissertation marks the end of a research project that I began nearly three and a half years ago, in October 2009. Looking back, I realize that it also represents the culmination of a quarter of a century spent in schools and universities, first in Lisse, and then in Leiden, Amsterdam and Ghent. During these years I have sat through more exams than I care to remember. The public defence of this dissertation will quite possibly be the last exam I will ever face.

It is customary at this point to acknowledge the help and support one has received from other people. First of all, I am grateful to my supervisors, Erik Weber and Jeroen van Bouwel; this dissertation could not have been written without their guidance. Over the years, there have been a lot of excellent teachers who, in different ways and at different times, have contributed to this achievement (in some cases without realizing it themselves). I cannot name them all, but among others, I have had the good fortune to have been taught by Jan van den Boom, Dé Steures, Wim Rosdorff, Jan Sleutels, Bert Bos, Herman Philipse, Gerard Visser, Theo Oudemans, Theo Meijering, Huib Looren de Jong, Hans Radder, and Henk de Regt. Also, many thanks to the members of the reading committee, to Prof. Dr. Freddy Mortier for chairing the examination committee, to Gitte Callaert for helping with the layout and design of this book, to the members of the Centre for Logic and Philosophy of Science for many fruitful hours of sharing views and exchanging ideas, and to Merel Lefevere in particular for helping me with all sorts of practical matters when I first moved to Belgium and took up work at Ghent University. A special thanks is extended to Dingmar van Eck, who was kind enough to proofread an earlier version of this dissertation and provide me with invaluable comments and suggestions.

To conclude on a more personal note, I am deeply indebted to my friends and family. Although they have had little to do with the contents of this dissertation, they have kept faith in me, which in any case is all that matters.

List of Figures

Table of Contents

Part 1

Background

Chapter 1
Introduction

1.1 Some general issues concerning scientific explanation

Of all the topics philosophers of science concern themselves with, few have received more attention than scientific explanation. The questions that have been raised in the literature are many: What is scientific explanation? What are its relata? What role does it play in the scientific process? What motivations do scientists have when they provide explanations? How can one distinguish between good and bad explanations?[1] Although the debate about scientific explanation undoubtedly has its roots in ancient philosophy (cf. Hankinson 1998), contemporary discussion of it really starts with the work of Hempel and Oppenheim in the mid-twentieth century, in particular with the so-called covering-law model of explanation (Hempel & Oppenheim 1948; Hempel 1965). Over the past decades, Hempel's work has been criticized by numerous philosophers of science; nevertheless, it continues to influence the present debate, most notably through the persistent idea that explaining something involves constructing arguments, a theme we shall often revisit in this dissertation.

[1] Despite the fact that many if not all of these issues have as of yet not been satisfactorily resolved, it is important to make one point clear from the outset: throughout most of this dissertation, the relata are assumed to be *statements* that, either qualitatively or quantitatively, describe events or states of affairs. This means that a sentence like 'Continental drift explains the shape of the continents' should be taken as shorthand for 'Statements describing continental drift explain statements describing the shape of the continents.' Of course, such statements come in many guises (they may be about particular events or general regularities, they may be intended as speculative or assertive, and they may be phrased in formal or informal jargon) and as we shall see, some explanations involve diagrams or flow-charts rather than statements; nevertheless, unless otherwise indicated, the relata are taken to be statements.

One of the reasons for all this attention is the intuitive notion we have that explanation somehow constitutes a central part of the scientific process. We feel that, in addition to merely describing the world, scientists are supposed to provide us with some form of *insight* or *understanding* (notions that themselves cry out for explication), and that it is the latter task that explanations are meant to accomplish.[2] Nevertheless, as we shall see, understanding is just one among many goals scientists may have in mind when seeking and constructing explanations.

When we try to answer questions about the nature, form and qualities (virtues and vices) of scientific explanation, we are engaged in a project that has its proper place in the philosophy of science. Of course, it is to be expected that any insightful account of explanation will inevitably draw on the resources of other philosophical domains, such as epistemology, logic, and metaphysics, but the focus will be on science itself. Philosophy of science in this sense involves a basic commitment to a weak form of naturalism, according to which one may only assume or postulate properties, entities or processes that are *compatible* with contemporary science (Sterelny 1990).[3] In the words of Michael Wheeler: "...if philosophy and natural science clash [...] then it is philosophy and not science that must give way" (Wheeler 2005 p. 5). On this account, metaphysics is subordinate to science. I will make some more specific comments on the methodological commitments of this dissertation in section 1.4.

In this dissertation, I will try to answer some of the questions formulated above from a *pragmatic point of view*, which involves (among other things) taking into account the notion that explanations are meant to achieve certain aims, and that these aims will not always be the same. Of course, this is only a very broad characterization of the phrase 'pragmatic', and shortly I will give a more detailed account of what I mean. First however, let me explain the motivation behind the decision to adopt this pragmatic point of view. This motivation is in part historical.

[2] This intuition, though very common, is not wholly uncontroversial. In the second half of the nineteenth century, there were some prominent theorists (like Ernst Mach and Pierre Duhem) who rejected explanation as a proper function of scientific theories.

[3] An example of a strong form of (ontological) naturalism would be the claim that any property, entity or process that acts as a cause of some physical effect must itself be physical.

1.2 The impasse between monism and pluralism as a motivation for a pragmatic approach

In the past, many philosophers have tried to construct theories of explanation. However, this type of endeavour is not without controversy. Is it possible to construct a single theory of explanation? Those who answer this question positively can be labelled monists, in that they try to capture all scientific explanations within the confines of a single, unified framework. Famous examples of monistic accounts include the already mentioned covering-law model, and Salmon's causal-mechanical account (1984). Yet monists have met with severe criticism over the past few decades. It seems that the hope of explicating a set of individually necessary and collectively sufficient conditions for explanation is just too ambitious a project: one can devise counterexamples to almost every definition. This is true for such venerable notions as causation, knowledge, identity and meaning, and it is no less true for explanation. Consequently, most of the major monistic accounts have had counterexamples levelled against them.

Given this situation, it seems a natural response to reject monism and adopt pluralism instead, and indeed, some monists have switched sides and joined the pluralist camp – Salmon himself may serve as an example (cf. Salmon 1998). Maybe scientific explanation is just a too heterogeneous, too dynamic affair to be captured in a single, across-the-board account. What counts as an explanation in one discipline may be very different from what counts as one in another discipline, and even within a certain discipline, many different forms of explanation may coexist. On this account, it seems that the monistic theories are often correct or insightful descriptions of *some* type of explanations, but not of explanation *tout court*, that is, not to the exclusion of other types. Sometimes, explanations are causal, sometimes not, sometimes they involve laws, sometimes not, sometimes they possess unifying power, sometimes not, and so on. As pluralists, we should allow all types of explanation their rightful place.

However, it is easy to see the pitfalls of this kind of reasoning. Although we might agree that we should allow different types of explanation to coexist, if it is impossible to put any general constraints on what counts as an explanation, then it is hard to see what insights a specifically philosophical analysis of it might yield. Adhering to an unspecified, *anything-goes* type of pluralism would relegate philosophy to a trivial, case-by-case description of an endless tide of concrete examples of explanation. Somehow, we must navigate between these two extremes. As pluralists, we must be tolerant of diversity, but not to the extent that our account becomes trivial. In effect, what we are looking for is a *sophisticated type* of pluralism, i.e. a pluralism that yields some positive insights into the practice of scientific explanation, without reverting back to monism.

This is exactly the aim of the *pragmatic approach to explanation* that has been developed over the past five years by members of the Centre for Logic and Philosophy of Science at Ghent University[4]. In chapter 4, I will give a more detailed account of this approach, but in a nutshell, it attempts to answer questions about the structure, scope and virtues of scientific explanations by taking into account the goals or *epistemic interests* these explanations are meant to achieve. It views explanations as answers to why-questions that are asked with specific purposes in mind. Within this framework, the different theories of explanation that have been developed by philosophers of science in the twentieth century are thought of as instruments, or tools in a toolbox, that are only used when the task at hand demands it.[5] In the past, this approach has been used to analyse debates on scientific explanation in such disciplines as history (Van Bouwel & Weber 2008a), social science (Van Bouwel & Weber 2008b), medicine (De Vreese, Weber & Van Bouwel 2010), psychology (Vanderbeeken & Weber 2002; Weber & Vanderbeeken 2005) and software engineering (De Winter 2010), among others.

1.3 Subject and aims of this dissertation

To summarize, the pragmatic approach analyses concrete examples of scientific explanation by taking into account the epistemic interests these explanations are meant to serve, and on the basis of this puts forward domain-specific descriptive and normative claims about explanation. To the extent that these claims generalize (e.g. apply more generally to scientific disciplines, or to certain species of explanation found in different disciplines) the approach aims to establish a sophisticated form of pluralism with respect to scientific explanation.

Typically, a dissertation starts with a statement of the problem the author wants to solve, and why he or she thinks it is important to do so. Mine is somewhat different in this respect: as I was brought in to contribute to a research project that was already in full swing, the problem was not of my choosing. Instead of building up a theory from scratch, this dissertation records the contributions that I have made to this on-going

[4] Although it has only been about five years since the term 'pragmatic approach' has been used explicitly to refer to a coherent research program (starting with Van Bouwel & Weber 2008a; 2008b), some of its theoretical concepts had already been developed about six years earlier (cf. Vanderbeeken & Weber 2002).

[5] With this in mind, one might say that rather than being yet another theory of explanation, the pragmatic approach is a meta-philosophical position on *how to study explanation*, while this meta-philosophical stance in turn is aimed at producing a pluralist theory of explanation at the object level – see section 4.1.

project while being enrolled as a PhD student at Ghent University. The general aim of my research project was to expand the scope of the pragmatic approach to explanation in cognitive science and biology. In practice, this meant addressing three lacunae in the approach – that is, it meant applying the approach to three related issues that, although discussed in the philosophical literature, had so far not received much attention from the researchers working within the framework of the pragmatic approach.

The first two contributions I made both relate to the fact that, generally speaking, the approach has so far been silent about the issue of *models in science*. As I applied the approach to debates about explanation in biology and the cognitive sciences, I had to study the practice of explaining by means of models, as this is widely recognized to constitute an integral part of the explanatory practice of biologists and cognitive scientists. This led me to focus on two issues.

The first issue is: how does explanation by means of a model relate to covering-law explanations? It seems that, following Machamer, Darden and Craver's seminal paper *Thinking About Mechanisms* (2000), the current trend among philosophers of science is to promote so-called mechanistic explanations as a superior alternative to covering-law explanations (Bechtel & Abrahamsen 2005; Craver 2007; Cummins 2000 – for an alternative view, see Leuridan 2010). Briefly, I argue that while the philosophical community is certainly right that until recently, mechanistic explanations have received too little attention, its out of hand dismissal of Hempel's model is premature. In fact, some forms of covering-law explanations (although not strictly conforming to Hempel's model, as I will explain) are still being used in cognitive science and biology, and indeed, are to a certain extent indispensable (see chapter 7). This is the first contribution I made to the pragmatic approach; it is presented in part II of this dissertation.

The second issue is: what epistemic interests can motivate model explanations? Mechanistic models are a contemporary alternative to *functional* models, as once described by Cummins (1975). By reflecting on the differences and commonalities between these two type of model explanations (again, a topic left largely untouched by the pragmatic approach[6]), I have identified three properties models can have: plausibility, richness and performance. Clarifying the relation between these properties is a prerequisite to answering questions about the *explanatory power* of model explanations in cognitive science and biology. Roughly speaking, functional models may lack implementational details, but they can be of explanatory value nonetheless, and so should not be dismissed, as they sometimes are (e.g. Craver 2006). The important factor is what you want to achieve with your model, and this determines whether, and to what

[6] Though not by the literature as such (e.g. Piccinini & Craver 2011).

extent, you want it to exhibit plausibility, richness, performance, or a combination of these properties. This is constitutes the second contribution, and it is presented in part III of this dissertation.

The third contribution is more theoretical in nature. Although philosophers working within the framework of the pragmatic approach have distinguished between different question-types to analyse particular explanations, they have not to a great extent reflected on the nature of, and the relations between, these different question-types. For example, whereas previously, the approach interpreted explanations as answers to why-questions, model explanations are often answers to *how*-questions. This has led me to compare the various properties these two question-types can have. Relatedly, there is the issue of *significance*. What makes a particular explanation-seeking question significant? Until recently, research conducted within the pragmatic approach took it as given that certain explanation-seeking questions are interesting while others are not. In the final part of this dissertation, I present some ideas as to what properties make an explanation-seeking question worth answering. The position emerging from these considerations is a sophisticated form of explanatory pluralism – robust pluralism, as I call it – which I describe and defend in part IV and in the concluding chapter of this dissertation.

Summarizing, the pragmatic approach is motivated by the impasse between monistic and pluralistic theories of scientific explanation, its aim is to develop a more sophisticated or robust form of pluralism, and my specific contributions to it are: a) to argue that, notwithstanding the importance of mechanistic explanations, certain types of covering-law explanations are still being used in cognitive science and biology, and are indeed indispensable, b) to argue that what properties a model exhibits (richness, plausibility and/or performance) is determined by ones epistemic interests, and finally c) to argue that by expanding the pragmatic approach to how-questions and introducing significance as a property of explanation-seeking questions and their answers, one can indeed arrive at a robust form of pluralism.

1.4 Methodology

In a sense, I have already explained the methodology of this dissertation, as the pragmatic approach is itself a method – namely, a method for studying scientific explanation. Nevertheless, there are some additional points worth noting.

The first point concerns the use of examples. The pragmatic approach is essentially a bottom-up approach. Instead of constructing a theoretical framework which is subsequently illustrated with examples, I proceed the other way round: I will analyse

examples to explicate the various concepts I need. That is, the stock of conceptual tools of the pragmatic approach is expanded as the need arises.

Second, there is the nature of the examples themselves. As I have said previously, the present dissertation will attempt to remain true to naturalism, in the sense that it follows an approach to explanation that is at least compatible with, and hopefully also an accurate description of, scientific practice. With this commitment in mind, I will illustrate the more substantial philosophical claims I make about scientific explanation with concrete examples taken from the scientific literature[7]. In other words: fictitious or everyday examples do not suffice. They will be employed now and then for heuristic purposes (to introduce a subject, or to explicate our intuitions regarding some issue), but to *justify* one's claims about science, it is necessary to provide examples of actual scientific explanations. Besides the general commitment to a weak form of naturalism that I have already mentioned, there is an additional reason not to rely on everyday examples: although explanation in ordinary, everyday situations may often be akin to explanation in a scientific context, there is no reason to suppose that this is always the case. At least, it does not follow without some additional argument. So while there are many situations in which mundane examples might do the job, the fact that there are situations in which they do not, means that we shall have to draw on the scientific literature itself.

Of course, following this methodology has consequences. From time to time, it will be necessary for me to provide some background information, introduce some jargon or make the reader familiar with certain theoretical assumptions of a particular field. However, the benefit is that the reader will be able to judge the relevance and descriptive accuracy of the pragmatic approach first hand, as it applies to concrete examples of scientific explanation. Of course, since my research has been directed at explanation in biology and the cognitive sciences, examples from these disciplines will loom large in this dissertation, but I will sometimes digress into neighbouring fields.

Finally, the aim of this thesis is not merely descriptive: explicating the interests explanations are meant to serve provides us with more than just the reason why a particular type of explanation has been chosen. It also contains a *normative* element, in that it allows us to make judgments about the strength and appropriateness of given explanations. Ultimately, such evaluative judgments can be used to develop guidelines or rules of thumb for future explanations. Thus, the normative dimension of the pragmatic approach comprises both evaluative and prescriptive elements.

---

[7] The only part where I have admittedly broken this rule is in chapter 13, where I discuss the various relations between why- and how-questions. This is not because I do not believe that the methodological principles should not apply to this particular issue, but simply because at the moment this dissertation was due for submission I had yet to find suitable examples.

1.5 Overview

This dissertation is divided into five parts. In the first part, I will present the necessary background theory, while in the second, I will present my findings on the relation between covering law and mechanistic explanations in cognitive science and biology. In part three, I will expand on three important properties of model explanations in these two disciplines, by comparing mechanistic explanations with Cummins-style functional explanations. In part four I will further develop the theoretical side of the pragmatic approach and the robust pluralism it leads to. Of special importance here are the relation between why- and how-questions and the issue of the significance of explanation-seeking questions. Finally, part five consists of one concluding chapter, in which I take stock and reflect on the prospects for future research. As these five parts are broken down into chapters, let me conclude this introduction by providing an overview of the ground we are going to cover chapter by chapter.

Although the starting point of this investigation is the pragmatic approach to explanation, in chapter 2 I will first briefly sketch the history of philosophic reflection on scientific explanation in the twentieth century. This is useful for two reasons. First, it will introduce many concepts and theories I will frequently refer back to throughout the succeeding chapters. Second, part of the original *motivation* for the pragmatic approach is enclosed in this history, particularly in the controversy between monistic and pluralistic accounts of explanation. In the second chapter, I will cover the basics of the covering law model (both the deductive-nomological and the inductive-statistical variants), the statistical relevance model, the causal-mechanical account, and unificationism, and discuss some of the problems that these models face.

In chapter 3 I will further analyse the dichotomy between monism and pluralism with respect to theories of scientific explanation. Furthermore, I will discuss in detail the so-called erotetic model of explanation, as developed by van Fraassen and subsequently used by other authors. There are two reasons for this discussion. First, the pragmatic approach to explanation shares with van Fraassen the basic idea that explanations are answers to certain types of questions, so understanding his account helps pave the way for chapter 4, where the pragmatic approach will be discussed. Second, although van Fraassen's account is pluralistic, it is pluralism *of a specific kind* – a kind that, as I will argue, fails to deliver. It is from the dialectics between monism and varieties of pluralism, that the particular kind of pluralism resulting from the pragmatic approach can best be understood. To this end, I will address various problems connected with van Fraassen's account and explain the concept of contrastive questions.

In chapter 4, I will present the pragmatic approach to scientific explanation, as it has been developed over the past years at Ghent University. I will introduce the concepts of epistemic interests and explanatory format. Explanations are always sought relative to

some goal, and as this goal partly determines the format of the explanation, it is imperative to consider the notion of epistemic interests in some detail. Finally, I will illustrate how the pragmatic approach works by drawing on an example from previously published material. This chapter concludes part I.

Chapter 5 begins with a short introduction of cognitive science. Although I will not really need the concept of functional models just yet, it will become important in part III, and the most natural way to introduce mechanistic models is through a comparison with functional models, which is the reason why this chapter includes such a comparison. In chapter 6 I will analyse two examples of explanation in dynamical cognitive science. The claim will be that these explanations are non-deductive causal CL explanations using a default rule. In chapter 7, I will argue that some explanations in biology are in fact a type of DN explanations, although they differ on a number of points from Hempel's original model. With this in mind, in chapter 8, after arguing that the conclusion of chapter 7 is controversial given the present state of the debate about biological explanation, I explore how Hempel's model could be adjusted to accord with this particular type DN explanations. This concludes part II

After introducing the topic of model explanations in chapter 9, I will focus on three properties models can have: plausibility, richness and performance. In chapter 10, I will argue that plausibility and richness vary independently, thus taking issue with the idea that the explanatory power of models is a function of their completeness as a description of a target mechanism. In chapter 11, I will explore how plausibility can be subordinate to performance in the context of artificial intelligence. This concludes part III.

In chapter 12 I further develop the pragmatic approach, showing how one can avoid all too liberal forms of pluralism by developing concrete guidelines by which to judge the significance of explanation-seeking questions and their answers. In chapter 13, I expand the pragmatic approach by applying it to how-questions, which, as the chapters on models explanations should have made clear, are important in their own right. This concludes part IV. Finally, in chapter 14, I will take stock, summarize the main results, and give some outlines for further research.

Chapter 2
Explanation in 20th century philosophy

2.1 Introduction

It is a time-honoured tradition to begin an overview of philosophical reflection on scientific explanation with the work of Hempel in the mid-twentieth century. Although the rationale behind it is by no means uncontested, this chapter will follow this conventional approach – not for the sake of tradition itself, but because it constitutes a clear framework in which to introduce a lot of material I shall need in later on in this dissertation. In the following sections, I will present what are usually taken to be the four most influential accounts of scientific explanation, together with their main difficulties. However, this chapter does not aim at completeness. As a lot of the more technical issues surrounding these models and their counterexamples will have little bearing on the main points I shall discuss later in this dissertation, I will mostly confine myself to giving the central ideas behind these theories, the exception being Hempel's covering-law model itself, which is treated in somewhat more detail. The four theories of scientific explanation I shall discuss are: Hempel's covering-law (CL) model, both in its deductive-nomological (DN) and inductive-statistical (IS) forms[8], Salmon's statistical-relevance model (SR), his subsequent causal-mechanical model (CM), and Kitcher's unificationist model (U). As van Fraassen's erotetic model of explanation is one of the main subjects of the next chapter, I will not consider it here. Being familiarized with

[8] Thus, following Hempel's example, I shall use the term 'covering-law model' to cover both the DN and IS varieties. CL designates the basic idea that both models have in common: explaining something is tantamount to showing it to be an instance of a more general regularity, whether this 'showing' is either achieved through deduction or induction. On both accounts, explanations are conceived as arguments.

these theories and some of the main objections that have been levelled against them will prepare us for the work that lies ahead.

2.2 The covering-law model

2.2.1 The deductive-nomological model

The basic intuition behind the DN model is that explaining an event involves providing a general regularity of which that event is a particular instance. Among members of the Vienna Circle in the 1930s, the idea was developed that the reference to a law is what marks a statement as a genuine scientific explanation, as opposed to the philosophical explanations involving reference to metaphysical agents such as entelechies or immaterial minds, prevalent in the German idealism that was still very much alive on the European continent at the time.[9] Thus, the two conditions that the DN model places on explanations are (i) that the explanans contains at least one empirical law (the nomological aspect of DN) and (ii) that the explanandum is derivable from the explanans (the deductive aspect). However, this is not enough: the first condition requires some fine-tuning.

It is not enough simply to state that the explanans must contain one empirical law, for two reasons. First, we cannot allow *just any old law* to do the job. The law has to be relevant in the sense that it is necessary for the derivation to take place. Remove it, and the argument should become invalid. Second, by themselves, laws do not state what will happen; they only state that if certain conditions are met, other things will happen. Therefore, if we want to explain the occurrence of a particular event, we should include a reference to the initial conditions that, together with the law, entail the explanandum. Thus, the explanans should include at least one statement that describes an empirical law and one describing some initial conditions.

Following Hempel and Oppenheim's (1948, p. 137) divisions, there are four conditions of adequacy for DN explanations (those familiar with Hempel and Oppenheim's original text will notice that I only quote the first sentence describing each condition):

[9] See Carnap 1966 for an account of the intellectual climate in Vienna that served as the bedrock for what would ultimately become Hempel's thesis.

Logical conditions of adequacy for DN explanations:

*R*1: The explanandum must be a logical consequence of the explanans.

*R*2: The explanans must contain general laws, and these must be essential for the derivation of the explanandum.

*R*3: The explanans must have empirical content, i.e. it must be capable, at least in principle, of test by experiment or observation.

Empirical condition of adequacy for DN explanations:

*R*4: The sentences in the explanans must be true.

According to the DN model then, explanations are deductively valid arguments conforming to the following schema (idem, p. 138):

$C_1, C_2, \ldots, C_k$
$L_1, L_2, \ldots, L_r$
==========
E

where $C_1, C_2, \ldots, C_k$ are statements describing the particular facts and initial conditions, $L_1, L_2, \ldots, L_r$ are statements describing the general laws, and E is a statement describing the explanandum. Arguments that comply with the logical conditions of adequacy R1-R3, and consequently adhere to this schema, are *potential* explanations. The fourth, empirical, condition of adequacy R4 goes beyond that: it requires the sentences making up the explanans to be true. It is only when this condition is met that we can speak, not merely of a potential, but of an *actual* explanation. In terms of the schema above, the DN model holds explanations to be *sound deductive arguments.* Although, as we will see, this model has met with severe criticism, the basic idea that explanations are (sound) arguments involving reference to some law or regularity is very persistent and continues to enjoy support to this day (e.g. Hausman 1998).

Two features of this model are important. First, the conditions of adequacy do not include the requirement that the explanans contains particular facts and initial conditions (the *C*-statements that make up the first line of the argument schema above), while these are necessary to explain the occurrence of particular events, as we have noted. The reason for this omission from the conditions of adequacy is that Hempel wants his DN model to apply not only to the explanation of singular facts, but also to

explanations of laws themselves. That is, he wants to accommodate cases in which one law or set of laws explain another law. This feature was important to contemporary philosophers, because it meant it could be used in the context of theory reduction, as indeed it was in Nagel's classic account or reductionism (1961).

Another important and well-known feature of the DN model is that there is no principled difference between explanation and prediction. As Hempel famously notes, these are really two sides of the same coin:

> Let us note here that the same formal analysis, including the four necessary conditions, applies to scientific prediction as well as to explanation. The difference between the two is of a pragmatic character. If E is given, i.e. if we know that the phenomenon described by E has occurred, and a suitable set of statements C_1, C_2, ... ,C_k, L_1, L_2, ... , L_r is provided afterwards, we speak of an explanation of the phenomenon in question. If the latter statements are given and E is derived prior to the occurrence of the phenomenon it describes, we speak of a prediction. It may be said, therefore, that an explanation is not fully adequate unless its explanans, if taken account of time, could have served as a basis for predicting the phenomenon under consideration (Hempel & Oppenheim 1948, p. 138).

Thus, the difference between them has to do with the epistemic order of discovery. This latter point is important, because it illustrates what according to Hempel is the major intellectual benefit explanations are meant to provide us with: "...the argument shows that, given the particular circumstances and the law in question, 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. 337). For Hempel, explanations are meant to provide us with knowledge of the world, where this knowledge is constituted by the expectability of the phenomenon we are explanatorily interested in.

2.2.2 The inductive-statistical model

As the DN model construes explanations as deductive arguments, it is obviously unable to accommodate explanations involving statistical laws. This is a serious shortcoming, as such explanations are common to many branches of scientific investigation, in particular to fields like genetics and medicine, where scientists frequently avail themselves of probabilistic laws. For example, the recovery of a patient from a streptococcus infection might be explained by the fact that the patient has taken

penicillin, together with a statistical law expressing the probability of patients recovering given that penicillin has been taken. Hempel recognized this problem, and so devised his inductive-statistical or IS model to remedy it (1962; 1965).[10]

Just as with the DN model, IS explanations are arguments that are required to have at least one law among their premises, the difference being that this law is now statistical in character, rather than a law of nature. Of course, this means that the conclusion no longer follows with deductive necessity. Instead, the premises must confer a high measure of probability upon the conclusion. Thus, although having taken penicillin is not a DN explanation of someone's recovery from streptococcus infection (there are some cases in which patients do not recover) it does presumably confer a probability of >0.5[11] and as such constitutes an IS explanation. Indeed, the strength of an IS explanation is proportionate to the inductive probability it confers on the explanandum.

Suppose then that we have a statistical law to the extent that *F*s have been found to be very likely *G*s with probability *r* (where we assume *r*>0.5), and a particular fact that *a* is *F*, then it is very likely that a is *G*. Schematically, the IS explanation would look like this:

Prob(*G*/*F*) = *r*
Fa
========== [*r*]
Ga

However, there is a catch. Unlike their DN counterparts, IS explanations notoriously suffer from the problem of ambiguity. This problem arises when we use statistical information about a class of events to make judgments concerning the probability of a single event; thus, it is a special case of the more general problem of selecting the proper reference class when applying a statistical generalization. Let us formulate the problem using Hempel's own example (1965 p. 395).

[10] In his (1965), Hempel also recognizes another type of statistical explanation, namely deductive-statistical or DS explanations, which involve deducing a narrower statistical correlation from a more general set of premises containing at least one more general statistical law. As the relation between the premises and the conclusion is one of deduction, the same conditions of adequacy which accompany DN explanations also apply to DS explanations.

[11] This assumption is in line with what Hempel called the *high probability requirement*, the idea that statistical laws explain a certain occurrence only if they confer a high probability on it.

Suppose we find (to our surprise) that November 27th in Stanford was a warm and sunny day, and we wish to explain this by means of an IS explanation, making use of a statistical law stating that the probability of any November day in Stanford being warm and sunny is 0.95. Let *n* stand for the particular day November 27th, *W* for the property of being warm and sunny, and *N* for the reference class, namely all November days in Stanford, and we get:

Explanation I

Prob (*W*/*N*) = 0.95
Nn
========== [0.95]
Wn

However, suppose we find that November 26th was cold and rainy, and that we have a statistical law saying that the probability that a cold and rainy day in Stanford is immediately succeeded by a warm and sunny day is 0.2. If we take *S* for our new reference class (namely immediate successors of cold and rainy days in Stanford), we can now construct an IS explanation to the effect that it is very likely that November 27th was in fact a cold and rainy day:

Explanation II

Prob (~*W*/*S*) = 0.8
Sn
========== [0.8]
~*Wn*

So, we have two perfectly good IS explanations (all premises of both are true), one explaining why November 27th was warm and sunny, and the other one explaining why the very same day was *not* warm and sunny. The problem stems, of course, from *n* being a member of both *N* and *S*. In other words, we need some way to choose the right or appropriate reference class.

Here, some philosophers might well point to pragmatic grounds for distinguishing between appropriate and inappropriate reference classes. After all, an explanation is usually sought *after* the explanandum has occurred, so in practice there would

presumably be no danger of someone accepting both explanations at once. For Hempel though, this will not do, because of the nomic expectability, or the logical isomorphism between prediction and explanation, which is at the heart of the CL model. Consequently, Hempel needs a more principled way of selecting the right reference class, and formulates what he calls the requirement of maximal specificity (RMS). Consider again our general IS schema:

Prob(G/F) = r
Fa
========== [r]
Ga

If we take S to be the conjunction of both premises (the statistical law and the particular fact) and K the set of statements that are accepted as true, RMS requires that if (S & K) implies that a is part of class F_1 and F_1 is a subclass of F, then (S & K) must imply some statement which explicates the probability r_1 of G in F_1, where $r_1 = r$ "...unless the probability statement just cited is simply a theorem of mathematical probability theory" (1965 p. 400). In effect, RMS requires that in constructing an IS explanation, we pick out the most specific (hence the name) reference class to which the explanandum belongs, i.e. the smallest reference class to which a is known to belong.

Thus, RMS constitutes an additional, empirical condition of adequacy when it comes to IS explanation (also, the first and second logical conditions are appropriately modified), so that we end up with the following five conditions:

Logical conditions of adequacy for IS explanations:

S1: The explanandum must follow from the explanans with high inductive probability.
S2: The explanans must contain at least one statistical law, and this must be essential for the derivation of the explanandum.
S3: The explanans must have empirical content; that is, it must be capable, at least in principle, of test by experiment or observation.

Empirical conditions of adequacy for IS explanations:

S4: The sentences in the explanans must be true.
S5: The statistical law in the explanans must satisfy the requirement of maximal specificity.

2.2.3 Problems with the CL model

Despite its status as the absolute classic, the CL model is not without its problems. In this section, I will consider three objections that have been levelled against it: the problems of *accidental generalisations*, *irrelevant premises*, and *asymmetry*.

First, the problem of *accidental generalisations*. This problem has to do with the notion of lawhood: The DN variant of the CL model must be able to distinguish between genuine laws and mere accidental generalisations. Thus, to use Hempel's own example, the statement that 'All members of the Greenbury School Board for 1964 are bald' is an accidental generalization in that if it is true, it is only contingently so, while the statement 'All gasses expand when heated under constant pressure' describes a genuine law. Although one can use the latter generalisation to explain why a particular volume of gas that is heated under pressure expands, one obviously cannot use the former generalisation to explain why a particular member of the Greenbury School Board for 1964 *n* is bald. Therefore, the DN model should be able to distinguish these two types of generalisations (accidental versus nomic).

In a similar vein, Salmon invites us to consider (1989, p. 15) the following two statements: 'No gold sphere has a mass greater than 100,000 kg' and 'No enriched uranium sphere has a mass greater than 100,000 kg'. Again, while the former statement only describes an accidental generalisation (it just so happens that no such gold sphere has been produced), the latter describes a law (because the critical mass of enriched uranium is only a few kilograms). It would appear that besides mere truth, modality is also important. Thus, mere subsumption is not sufficient: if we cannot make the distinction between accidental generalisations and genuine laws, it is not clear how the DN model can avoid such counterexamples.[12]

Second, the problem of *irrelevant premises* raises issues having to do with the notion of deduction. Specifically, it is always possible to add some superfluous yet true premise to a sound argument, and the result will still be a sound argument. However, such arguments intuitively do not constitute explanations. Thus, as Kyburg's (1965) classic example goes, we do not want to count

[12] As should be clear from the Greenbury School example, Hempel himself acknowledged this difficulty, and in his (1965) considered a number of candidate notions for lawhood, none of which he finds satisfactory. Now one might object that it is unfair on Hempel to count the lack of such a notion as an argument against his model, since the issue of what constitutes lawhood is a controversial one and has, even to this day, not been solved in a way that has won universal acceptance among philosophers of science. However, as lawhood is so central to the DN account, it seems philosophically unsatisfying to treat the notion as a primitive.

All samples of hexed salt dissolve in water
I have hexed this sample of salt
==========
This sample of salt dissolves in water

as an explanation, even though it satisfies Hempel's criteria, because salt dissolves in water anyway, regardless of it being hexed or not. Note that this objection can be adapted to provide a counterexample to the IS model as well: one simply has to substitute the law 'All samples of hexed salt dissolve in water' with a statistical generalization, such as 'Samples of hexed salt have the probability of 0.95 of dissolving in water.' Although this statistical generalization can be used to predict that a particular sample of hexed salt dissolves in water, it does not explain it.

Another classic is due to Salmon (1971b p. 34). Consider the argument: 'All males who take birth control pills regularly fail to get pregnant. Jones is a male who regularly takes birth control pills. Hence, Jones fails to get pregnant.' Again, the premises are true and the inference is valid, but of course the first premise is superfluous as males do not get pregnant anyway. Hence, the problem of irrelevant premises constitutes an argument against the DN model being sufficient.

Finally, the problem of *asymmetry* of explanation stems from the fact that while arguments can be reversed, explanations cannot.[13] Again we are faced with counterexamples. Consider:

Question: Why does the flagpole have a shadow of 10 metres long? *Answer*: The flagpole is 10 metres high. The sun is at a 45° angle above the earth. Light moves in a straight line, so we can derive that the shadow is 10 metres long.

Question: Why is the flagpole 10 metres high? *Answer*: Its shadow measures 10 metres in length. The sun is at a 45° angle above the earth. Light moves in a straight line, so we can derive that the flagpole is 10 metres high.

Both these arguments fit the Hempelian model; intuitively however, it is clear that only one constitutes a genuine explanation. Of course, from the length of the flagpole, the Pythagorean Theorem together with information about the behaviour of light and the elevation of the sun, we can derive the length of the shadow, and this results in a successful DN explanation. As far as deduction is concerned however, from the very same Pythagorean Theorem and the very same information about the physics of light, in conjunction with the length of the shadow, we can derive the length of the flagpole.

[13] The remainder of this section has, with some slight adaptations, been published in Gervais & Weber (2011).

Intuitively, this is not a sound explanation, yet from a logical point of view, the deductive inference is perfectly valid. If you know two of the variables, you can work out the third. Thus, the asymmetry problem constitutes a third argument why the DN model is not sufficient.[14]

2.2.4 An obvious solution

It should be fairly obvious that the counterexamples described in the previous section all point to the conclusion that something is missing from the CL model: an extra requirement to the effect that the premises of an argument must refer to the *cause* of the explanans should be added. This approach is championed by (among others) Hausman, who introduces the criterion of independent alterability:

> *Independent alterability* – For every pair of variables, X and Y, whose values are specified in a derivation, if the value of X were changed by intervention, then the value of Y would be unchanged (1998 p. 167).

It is easy to see how this criterion remedies some of the shortcomings of the CL model. Membership of the Greenbury School Board for 1964 is not a cause of *n* being bald. It is the Jones' being a man, not his taking birth control pills, that cause him not to get pregnant. It is the height of the flagpole that causes the length of the shadow, and not the other way round. Thus, by adding this extra requirement, Hausman's adaptation of the CL model manages to avoid the counterexamples mentioned above, while still maintaining that explanations are arguments (they are just a special kind of arguments).[15,16]

---

[14] Nevertheless, Hempel himself maintained that explanations are symmetrical. For him, the flagpole example only shows that our intuitions about explanation are mistaken.

[15] However, it may be that even with the causal requirement added, the CL model still faces problems: counterexamples have been constructed to the effect that arguments complying with the extra causal condition need not be explanatory (see Ruben 1990 for an excellent discussion). Exploring this matter any further would take us too far afield for the purposes of this introductory chapter.

[16] Some philosophers, while agreeing with Hausman's diagnosis that the CL model lacks a notion of cause, disagree with him that simply adding a causal requirement to the existing argument structure is sufficient: they reject the very idea that explanations are arguments altogether, and hence, that they include anything like causal derivations (apart from Salmon, examples of such non-derivationist causalists are Nancy Cartwright and Paul Humphreys). Of course, by jettisoning derivability as a requirement for explanation, they reject Hempel's basic insight of nomic subsumption. Unlike Hausman's position then, their arguments hardly constitute a repair of the CL model.

2.3 The statistical relevance model

In line with the arguments given in section 2.2.3 and 2.2.4, many current philosophers of science recognize the need to include some notion of causality in their model of explanation.[17] One way to characterize this notion is by interpreting it as conditional or statistical dependency relationships.[18] A very influential account of explanation that takes such a statistical dependency notion as its starting point is Salmon's Statistical Relevance (SR) model (Salmon, 1971a). The basic idea here is that only those events or properties that are statistically relevant to another event or property are explanatory, while those that are statistically irrelevant are not. Of course, this requires the notion of statistical relevance to be spelled out more fully. The definition Salmon gives is that given a class or population A, C will be statistically relevant to B if and only if the probability of B conditional on A and C is different from the probability of B conditional on A alone. More formally:

$$C \; sr \; B \leftrightarrow \text{Prob}(B \mid A.C) \neq \text{Prob}(B \mid A)$$

It should be clear that a definition like this avoids some of the awkward counterexamples raised in section 2.2.3. Consider the hexed salt example. Here, the population A includes both hexed and non-hexed examples of salt. Filling in the other variables in the definition, we get: the probability of a sample of salt dissolving, conditional on it being a member of population A and it being hexed is 1, while the probability of it dissolving conditional on it being a member of A alone is also 1. As the definition requires these values to not be equal, we can conclude that being hexed is not statistically relevant for dissolving, and hence is not explanatory. Similarly, in the example of the birth control pills, the probability of Jones getting pregnant while taking birth control pills is equal to the probability of him getting pregnant and not taking birth control pills, and hence his taking birth control pills is statistically irrelevant to

[17] Although Hempel forgoes any appeal to causality (as he is unwilling to either construct a detailed account of his own, or to simply treat causation as a primitive), one could argue that the notion of causality inherent in the covering law model takes the form of a Humean regularity account, where a causal claim implies nothing more than that there exists some regularity linking cause with effect. With this in mind, one could say that the arguments from sections 2.2.3 and 2.2.4 show that, when a is regularly followed by b, the regularity account is not sufficient to establish the truth of a claim asserting a causal link between a and b.

[18] Nevertheless, it is doubtful whether causal relations can be captured under the notion of statistical relevance relationships, as the former are underdetermined by the latter (cf. Suppes 1970; Spirtes, Glymour & Scheines 1993/2000).

him getting pregnant or not. Or, if we take *A* to include both males and females, we can conclude that taking birth control pills is only statistically relevant to the female members of A.

As the SR model has only very limited relevance for the main themes developed in this dissertation, I will forgo some of the more technical issues surrounding it and instead note two key differences with Hempel's IS model. First, although both centre on statistical relations, SR rejects the notion that explanations are (inductive) arguments. For Salmon, what distinguishes good arguments from bad arguments and good explanations from bad explanations is simply not the same. Second, unlike the IS model, which, as the reader will recall, required that the explanandum must follow from the explanans with high inductive probability (the first logical condition of adequacy) the definition of SR does not include a similar requirement. Low probabilities are also explanatory. This is both good and bad news. On the upside, it means that certain low-probability effects are nevertheless explained by their causes, just as our intuition would have them be. For example, although the chances of contracting the West Nile virus from a single mosquito bite are presumably lower than 0.5, nevertheless, if someone does contract the virus by a single mosquito bite, most of us would call that bite explanatory relevant. In contrast to IS, SR accommodates such low-probability explanations. On the downside, this accommodation of low-probability explanations does have the intuitively unsavoury consequence that a single explanans can explain both a state of affairs *S* and other states of affairs that are actually inconsistent with S, such as ~S. In the example just given, both individual 'contractings' and 'non-contractings' (pardon the English) of the West Nile virus are explained by the same statistical relation.

2.4 Cartwright on positive causal factors

As we have seen in the preceding sections, introducing some kind of causal requirement is an obvious way to avoid some of the counterexamples Hempel's IS model suffers from. Nancy Cartwright agrees with Hausman that causality is necessary for explanation, yet she also agrees with Salmon that although some explanantia confer only low-probability on their explananda, they are nevertheless explanatory: a single mosquito bite can explain someone contracting the West Nile virus. However, Cartwright draws a different conclusion. For her, the possibility of low-probability explanations shows that explanations are not arguments. Being bitten by a mosquito is a *positive causal factor* in contracting the West Nile virus. For Cartwright, the explanans must contain one or more causes and increase the probability of the explanandum. In this way, she hopes to

avoid the awkward consequence of Salmon's SR model mentioned at the end of the previous section, namely that the explanans also explains other states of events that are inconsistent with the original explanandum. Thus, according to Cartwright, we do not explain why someone did not contract the West Nile virus by pointing out that the individual was bitten by a mosquito, because a mosquito bite is not a positive causal factor for not contracting the virus. To give her own example:

> I consider eradicating the poison oak at the bottom of my garden by spraying it with defoliant. The can of defoliant claims that it is 90 per cent effective; that is, the probability of a plant's dying given that it is sprayed is .9, and the probability of it surviving is .1. Here (...) only the probable outcome, and not the improbable, is explained by the spraying. One can explain why some plants died by remarking that they were sprayed with a powerful defoliant; but this will not explain why some survive (1983 p. 28).

2.5 The causal-mechanical model

Having abandoned SR in the face of severe difficulties[19], Salmon later proposed another causal approach to explanation: the causal mechanical (CM) model (Salmon 1984). The CM model tries to capture explanation in terms of *physical causal processes*.[20] These physical processes are often typified in robust terms: they consist of objects pulling and shoving, attracting or colliding with each other (baseballs shattering windows is a popular example). In general, the sort of physical processes Salmon envisages are capable of transmitting a mark (note the word *capable*: it is not necessary for them to *actually* transmit a mark). Mark transmission is the ability of a process to produce a change that 'carries over' to another spatiotemporal location, as happens when a baseball shatters a window. In this sense, the mark transmission is contiguous, as it persists through changes in time and place.

For Salmon, mark transmission is what distinguishes genuinely causal processes from mere pseudo-processes, such as the familiar shadow of a moving car. To borrow Salmon's own example (1984 pp. 141-142), suppose we have a large circular building, with a white spotlight in the centre, mounted on a revolving platform. If the spotlight is

[19] For a clear and concise overview of Salmon's reasons for abandoning SR, see his 1990.

[20] Having defined the relata as such, the CM model could be viewed as a species of, or at least being compatible with, a process theory of causation like the one developed by Dowe (2000).

switched on and the building is darkened, we will see a moving spot of white light moving along the wall, and if we place a red filter in front of the spotlight, this will change the colour of the moving spot. Hence, we are dealing with a genuine causal process, in that it is capable of transmitting a mark. Let us now suppose that we place the filter, not directly in front of the spotlight, but on the wall, the spot will again change its colour. However, as soon as the spot moves from its position, it will be white again. The only way we can make the change persisting, is by running along the wall, holding the filter in front of the spot. The moral should be clear: although a pseudo-process can make a mark, it cannot transmit it. The mark can only be maintained by a continuous effort. In this sense, Salmon's interpretation of pseudo-processes is strangely reminiscent of the seventeenth century doctrine of occasionalism.

In short, causal processes are physical processes that are capable of transmitting a mark. The mark transmission itself is a spatiotemporal intersection between two processes, such that the structure of both processes is changed. Thus, according to Salmon, explaining some event means explicating the causal processes and interactions that bring about the event (the *etiological* aspect of explanation), and describing the processes that comprise the event itself (the *constitutive* aspect of explanation). Such explanations show how the explananda "...fit into a causal nexus" (1984 p. 9).

As with CL and SR, many objections have been raised against CM. Some of these are general problems that plague all process theories of causality, for example, the difficulty in accounting for causation by omission and prevention (see Dowe 2003, Schaffer 2003 and Craver 2007 for a discussion). Kitcher has cast doubt on (among other elements of CM) the idea that pseudo-processes cannot transmit marks (1989 p. 463, the 'problem of derivative marks'). Still others have argued that Salmon's analysis of causality is circular (Kitcher 1985, Mellor 1988). Finally, the case has been made that CM has trouble in excluding irrelevant causal processes as explanations (Hitchcock 1995). Suppose that the baseball shattering the window was hit with a freshly painted bat, so that upon impact, the bat transfers some of its paint to the ball. If we want to explain the broken window, we want to refer to the relevant causal process, namely the one involving the mass and velocity of the bat and the ball, rather than this smudge of paint. Nevertheless, the latter is clearly a transmitted mark in Salmon's sense, and so constitutes a genuine causal process. Why then is the smudge of paint explanatory irrelevant? Of course, local and more extensive repairs can be made to CM to avoid some of these difficulties.[21] However, the lesson should be that by itself, mark transmission is neither necessary nor sufficient for processes to be causal, let alone be of explanatory value.

[21] See Salmon 1994, where he tries to avoid some of the counterexamples given in Kitcher 1989.

2.6 Unificationism[22]

According to Kitcher (1989), explanation is aimed at what he calls *unification* (U). At an intuitive level, unification is about bringing the many back to the few, about systemizing our knowledge and bringing apparently disparate phenomena together into a coherent framework.[23] In an oft quoted passage, Kitcher captures the core of his theory[24]:

> *Science advances our understanding of nature by showing us how to derive descriptions of many phenomena, using the same patterns of derivation again and again, and, in demonstrating this, it teaches us how to reduce the number of types of facts we have to accept as ultimate (or brute)* (1989, p. 432 italics in original).

Note that in this quote, Kitcher employs the notions 'derivation', and 'derivation pattern', elements which are reminiscent of CL. Indeed, it would be wrong to say that U serves as a competitor for CL, and Kitcher himself saw it as the implicit or unofficial model underlying CL (1981 p. 508). Nevertheless, while Kitcher shares Hempel's conviction that all explanations are valid arguments, he does not hold the converse, i.e. that all valid arguments are explanations.

In my view, Kitcher's position is undeniably appealing. It seems uncontroversial that unification, as understood in the quote above, is both an important goal of science, and a source of scientific progress. Kitcher's idea is that unification involves deriving descriptions of diverse phenomena from as few argument patterns as possible. As Kitcher retains Hempel's notion that to explain is to produce an argument the conclusion of which is a description of the explanandum (1989: 431), to unify is to show that various covering law arguments instantiate a single argument pattern. So what constitutes an argument pattern?

Kitcher introduces several technical notions (1989 p. 432). First, there are *schematic sentences*; ordinary propositions of which some (though not necessarily all) non-logical words are replaced by dummy letters. To borrow Kitcher's own example, a sentence like

---

[22] Here, the postponed treatment of van Fraassen's model of explanation (see section 2.1) has the admittedly awkward consequence of bereaving Kitcher's unificationism of some of its initial appeal. As Kitcher himself emphasized (1989 p. 415), one motivation for unificationism is to avoid the danger of trivialization that a purely pragmatic theory like the one offered by van Fraassen (with its heavy focus on the psychological factors like background knowledge) faces. However, these concerns regarding van Fraassen's account will be spelled out in detail in the next chapter.

[23] Part of this section is taken from Gervais & Wieland (2012).

[24] A similar idea was already developed by Friedman (1974). Kitcher readily acknowledges the intellectual debt he owes to his predecessor (1989 p. 431).

'Organisms homozygous for the sickling allele develop sickle-cell anaemia' can generate schematic sentences like 'Organisms homozygous for *A* develop *P*' and 'For all *x*, if *x* is *O* and *A* then *x* is *P*'. Exactly what these dummy letters will stand for in a given situation is then determined by filling instructions. These rules give direction to the process of substitution. Thus, in the case of the first schematic sentence, the instruction might be to replace *A* with the name of a specific allele and *P* with the name of a phenotypic trait. A *schematic argument* is a sequence of schematic sentences, while a *classification* of such an argument describes its inferential rules. Finally, combining a schematic argument, a set of filling instructions for each term, and the appropriate classification, gives us a *general argument pattern*.

To see how all this works in practice, let us briefly consider one of Kitcher's examples: the explanation of homologous traits in organisms. If we take *P* as the trait, *G* and *G** as groups of organisms, and *S* as the species (the filling instructions), we can construct the following derivation pattern (1989 p. 443):

(1) *G* and *G** are descendant from a common ancestral species *S*.

(2) Almost all organisms in *S* had property *P*.

(3) P was stable in the lineage leading from *S* to *G*; that is, if *S* was ancestral to S_n, and S_n was immediately ancestral to S_{n+1} and S_{n+1} was ancestral to *G*, then if *P* was prevalent in S_n, almost all members of S_{n+1} were the offspring of parents, both of whom had *P*.

(4) P was stable in the lineage leading from *S* to *G**

(5) *P* is heritable; that is, almost all offspring of parents both of whom have *P* will have *P*.

(6) Almost all members of *G* have *P* and almost all members of *G** have *P*.

The classification of this argument is that (1) to (5) are premises, and (6) is the conclusion derived from the premises by means of mathematical induction.

However, it takes more for a concrete argument to be an explanation than merely being an instantiation of this general pattern. The argument must instantiate an argument pattern that belongs to a set of derivation patterns: *the explanatory store E*(*K*); i.e. the set of patterns that maximally unify *K*. An important property of a general argument pattern is its *stringency*: the more constraints it puts on the arguments instantiating it, the more stringent it is. On this account, the stringency of an argument pattern allows us to see the common patterns among groups of derivations. Thus, the idea summarized in the quote at the beginning of this section is that we explain by

deriving descriptions of multiple and disparate phenomena using as few and as stringent general argument patterns. Of course, this makes unification a gradual notion. In turn, it allows us to make normative judgments about explanations: argument patterns with greater unifying power constitute better explanations.

I have already briefly commented on some important similarities and differences between U and CL. However, Kitcher's model can be viewed as superior because it is able to accommodate some of the counterexamples against Hempel's model. For example, the problem of irrelevant premises can be elegantly solved by U. Recall the hexed salt example. On Kitcher's account, an explanation of why a sample of salt dissolves that appeals to the sample being hexed, would be defective when compared with an explanation that only cites the generalization that all samples of salt dissolve in water. This is because the explanatory store containing only the former but not the latter would be less unifying than the converse one, as it does not permit the derivation of facts about unhexed samples. Adding generalizations about the solubility of hexed samples of salt to an explanatory store already containing generalizations about the solubility of salt in general, would only needlessly increase the number of argument patterns. Similarly, U is able to avoid problems associated with the asymmetry of explanation. If we take the example of the flagpole, the argument we want to avoid takes the following form:

L height = shadow·tan(angle).
C_1 A shadow is x meters.
C_2 The angle of elevation of the sun is y°.
==========
E The height of the object which casts the shadow is x·tan(y) meters.

So how does U rule out this argument as an explanation? According to Kitcher, the flagpole's shadow does not explain the flagpole's height as there is an alternative derivation of the latter which instantiates an argument pattern with greater unifying power. Here Kitcher introduces the notion of an origin-and-development pattern, which is a "...general pattern of tracing the present dimensions to the conditions in which the object originated and modifications that it has since undergone" (1989 p. 485). Consider the fact that unlighted flagpoles also have heights; in this case, if we wanted to explain the height by the shadow, we would need a different explanation of the height (i.e. another argument pattern) in case it is dark. More generally, we would end up with different explanations for the height of lighted and unlighted things, and this is not the case for the derivation of the flagpole's height which invokes an origin-and-

development argument pattern (as Kitcher suggests), rather than shadows. Origin-and-development argument patterns refer to whatever process produced and maintains the object, and so capture a broader set of explanatory situations. In the case of the flagpole, such an origin-and-development pattern will presumably refer to the manufacturing process of the flagpole, which explains its height regardless of the presence or absence of shadows.

Although it views explanations as deductive arguments, U is able to avoid some of the problems associated with DN. However, U has problems of its own. Let me conclude this section by briefly considering what (in my view) is the most urgent one: the problem of spurious unification.[25, 26] Roughly speaking, the objection is that according to U, any fact *F* can explained by a derivation from *F* itself. Take the argument pattern:

α
========
α

where this pattern is accompanied by the filling instruction: substitute α for any statement we accept. On the face of it, this argument pattern poses no problem for U, since it lacks any stringency. However, it is possible to make such a pattern more stringent by introducing restrictions. Suppose that one pattern we do not wish to exclude (e.g. the one about homologous traits in organisms above) is used to formulate conclusions adhering to condition *C*. In that case, one may propose filling instructions to the effect that α is substituted by a sentence which also adheres to *C* (as opposed to any accepted sentence). It seems that the argument pattern above is stringent, yet remains spurious. Obviously, proponents of U will want to avoid this.

[25] This problem was raised by Kitcher himself (1981 p. 526-529), although it is curiously absent from his 1989. For a critical assessment of Kitcher's attempt to overcome the problem, see Gijsbers 2007.

[26] For a detailed overview of the problems surrounding Kitcher's view, see Woodward 2003, pp. 358-373.

Chapter 3
Monism and varieties of pluralism

3.1 Introduction: monism and pluralism[27]

As the CL model is beset by the various problems and counterexamples we have been discussing, many philosophers reacted to this situation by constructing alternative accounts of explanation – in the previous chapter, we have encountered some of the most influential ones. What all these alternatives have in common with CL, however, is that they are *general*, *monist* theories. That is, they seek to *explicate* the concept of explanation, as it is used across the whole of science, or at least within some particular scientific discipline. In this sense, they stay within the general philosophical framework defined by Carnap (1950), for whom explication was, roughly speaking, the act of finding a scientifically fruitful alternative concept (explicatum) for the target concept (explicandum), and exactly formulating the rules of use for that new concept.

However, as the many objections and counterexamples given in the previous chapter should make plain, this project seems too ambitious. For Carnap, the explicatum has to be sufficiently similar to the explicandum, so that in most cases they may be used interchangeably. Even if we forgo the necessity of close similarity and allow, as Carnap himself did[28], considerable differences between explicatum and explicandum, it should

[27] Part of this section is taken from Gervais (2012c).

[28] The reason Carnap thought close similarity is not required is that other considerations, in particular what he calls fruitfulness, i.e. usefulness of a concept in formulating many universal statements (1950 p. 7), may lead scientists to replace an explicandum with an explicatum that is much narrower in meaning. To give his own example, the scientific concept of 'Piscis' (explicatum) does not even approximately coincide with the prescientific concept 'fish' (explicandum) as the former excludes whales and seals, while the latter (taken to refer to any sea dwelling creature) does not; nevertheless "...zoölogists found that the animals to which the concept Fish applies (...) have by far not as many other properties in common as the animals which live in

be obvious that there is no substantial overlap between, for example, the concept of scientific explanation (the explicandum) and Hempel's DN model (the explicandum). In short, all attempts discussed in the previous chapter to explicate the concept of explanation in Carnap's sense have failed, let alone the much more stringent procedure of providing a set of necessary conditions that have to be collectively sufficient.

Indeed, reflecting on the history and practice of science, the lesson seems to be that scientific explanation is simply a too dynamic and heterogeneous affair to be captured within the framework of a single, unified theory. As will be evident from the examples considered in parts II and III of this dissertation, sometimes explanations are causal, sometimes not, sometimes they invoke laws, sometimes not, sometimes they take the form of a model, sometimes not. What counts as a good explanatory strategy in one field does not necessarily carry over to other fields – and within the disciplines themselves, different types of explanations can co-exist. One could say that the theories considered in the previous chapter all highlight one or more important intuitions we have about explanation, often to the exclusion of others. A Carnap-style explication of scientific explanation is only to be had on a *local* level, since the possibility of formulating the rules of *all* uses of that concept is simply not on the cards.

However, maybe we should simply embrace this situation, that is, maybe we should abandon monism in favour of *pluralism*. Explanatory pluralism is roughly the view that different explanation types coexist. Unlike classic monistic models of explanation, explanatory pluralism maintains that one cannot rule out any type of explanation, and that it is not possible to make general evaluative judgments about different types of explanation. To put it more concisely, I take explanatory pluralists to subscribe to the following two theses:[29]

A) There are no general exclusion rules with respect to scientific explanations.

B) There are no general preference rules with respect to scientific explanations.

For example, one cannot in general rule out CL explanations in favour of CM explanations, or (truthfully) claim that CM explanations are always better than CL explanations. Both types of explanation may be used by scientists, and any evaluative

water, are cold-blooded vertebrates, and have gills throughout life. Hence the concept Piscis (...) allows more general statements than any concept defined so as to be more similar to fish; and this is what makes the concept Piscis more fruitful" (1950 p. 6).

[29] The following two theses are adapted from Van Bouwel & Weber 2008a.

claim can only be made relative to the context of the individual explanation. In this way, explanatory pluralists seek to avoid the counterexamples that have been levelled against the old monistic models, and make room for the dynamic and diverse nature of scientific explanation. Let us call the conjunction of these theses *minimal pluralism*: A and B are the basic claims pluralists commit themselves to, although of course they may hold additional viewpoints (we shall take a closer look at possible varieties of pluralism in section 3.4). One important pluralist in this sense is van Fraassen, to whom I will now turn.

3.2 The erotetic model of explanation[30]

According to van Fraassen (1980), explanations are answers to why-questions – this is the upshot of his so-called erotetic (from the Greek *erõtésis*, that which concerns questions) model of explanation. For van Fraassen, why-questions typically have the form:

> "Why A rather than B"?

Such questions have three important features: the *topic* (in this case A) which is taken to be true, the *contrast class* (in this case B), a class of propositions which includes the topic, and a *relevance relation*. The topic represents the fact in need of explanation. A why-question thus contrasts the topic with a member of the contrast class, although this contrast is not always explicitly mentioned. For example, we might ask "Why does John paint a portrait of the Queen?". Stated like this, the question is ambiguous. We might want to know why John paints a portrait of the *Queen* rather than a landscape. Or perhaps our interest is still different, and we would like to know why *John* paints a portrait of the Queen, rather than someone else (perhaps it is public knowledge that John has never finished art school). To remove this kind of ambiguity, we must choose a particular proposition from the contrast class to, well, contrast with the topic: this is called the *foil*.[31] Thus, in a question such as: "Why does John paint a portrait of the queen rather than an accomplished painter?" we have explicated the foil as well as the topic.

[30] Parts of this section is taken from Weber, Gervais & van Bouwel (2013).

[31] The term 'foil' has become standard, but it is not used by van Fraassen (at least not in his 1980). Sometimes, the term is used as equivalent for van Fraassen's contrast class. However, they are not the same, as the

Yet even with these explications in place, there remains a danger of ambiguity: why-questions can be formulated with identical topics and foils, yet still demand different answers. Suppose that, at the end of a trial conducted somewhere in the U.S., a defendant is sentenced to a year imprisonment. We might ask: "Why is the defendant sentenced to a year imprisonment, rather than acquitted?" For someone who is unfamiliar with the American legal system, the answer "Because the jury found him guilty" can be informative. Yet for another person well acquainted with American law, this answer is entirely unsatisfactory. For him, the answer should probably say something as to *how* the jury came to their decision. It seems that explicating the foil is not always sufficient to remove the ambiguity from a why-question. In these cases, the third element, the relevance relation, has to be specified. In the example of the defendant, two relevance relations come into play. In the case in which the explainee is unfamiliar with the American legal system, he or she expects causal factors as answers (because he or she does not know them), in the case in which the explainee does possess knowledge of the American legal system, he or she expects a description of the (unknown) causal mechanism linking (known) causal factors to the effect to be explained.

To put it more formally, a why-question Q consists of and is determined by three elements (1980 p. 143):

> The topic P_k
> The contrast-class $X = \{P_1, \ldots, P_k, \ldots\}$
> The relevance relation R

With these elements in place, abstract why-questions like "Why A rather than B?" are specified as follows:

$$Q = \langle P_k, X, R \rangle$$

where an answer A is relevant to Q if it stands in R to the couple $\langle P_k, X \rangle$. P_k is true in contrast with the rest of X because of A. So van Fraassen's view on what counts as a legitimate why-question and what as a legitimate answer can respectively be stated as follows:

---

contrast class is not a single proposition but a *set* of propositions. Furthermore, while the foil is false according to van Fraassen, the contrast class contains one true proposition, namely the topic (1980 p. 142).

α) It is worthwhile to attempt to answer the contrastive question "Why X rather than Y?" if and only if (a) X is true and (b) Y is false.

β) An answer to a contrastive why-question is an adequate explanation if and only if (a) the question is about a true contrast and (b) the answer is true and stands in the contextually determined relevance relation R to X.

We can now appreciate that van Fraassen's account is thoroughly pragmatic. The contrast class chosen in Q, and the relevance relation R determining the relevance for A vis-à-vis Q, are determined by pragmatic factors. Ultimately, it is the context that determines what counts as a genuine explanation-seeking question and what counts as a genuine answer/explanation – other than that, there is no principled constraint on R:

> So scientific explanation is not (pure) science but an application of science. It is a use of science to satisfy certain of our desires; and these desires are quite specific in a specific context, but they are always desires for descriptive information [...] The exact content of the desire, and the evaluation of how well it is satisfied, varies from context to context. It is not a single desire, the same in all cases, for a special sort of thing, but rather, in each case, a different desire for something of a quite familiar sort. (1980, p. 156)

As we will see in the next section, this liberalism invites a number of problems. To conclude this section however, it might be prudent to point out three claims van Fraassen apparently commits himself to: first, that all explanation-seeking questions are why-questions (1980 p. 134), second, that all explanation-seeking questions are contrastive, and third, that all foils are false.[32] In the course of this dissertation, I will reject all three of these claims, marking the place where I do so.

3.3 Green cheese, red herrings, and dead Kennedys

As has been pointed out by Kitcher and Salmon (1987), van Fraassen does not put any formal constraint on *R*. This means that the model is in danger of becoming trivial. It suffers from what I shall cautiously term excessive liberalism.[33] Critics claim that the

[32] Van Fraassen was not alone in this last assertion (cf. Garfinkel 1981 p. 40; Ruben 1987; Temple 1988 p. 144).

[33] At least, the model as presented in his (1980) does. I am not aware of any restrictions van Fraassen has made in his subsequent work.

erotetic model allows too much choice when it comes to the foil and relevance relation, and they have constructed numerous examples of bogus why-questions and answers that cannot be excluded by van Fraassen's original account. Two of such counterexamples are the so-called 'green cheese' and 'red herring' problems (the labels are from Risjord 2000a).

First, the green cheese problem. This is a problem of excluding uninteresting foils. Every state of affairs that we might be explanatorily interested in can always be contrasted with countless other states of affairs that are not the case, but the vast majority of which lead to questions that (a) most people would regard as strange and (b) scientists would not consider interesting. For instance, the topic "The level of unemployment in Belgium is 7.4 %[34]" might be contrasted with an infinite set of states of affairs which are not the case, but the most of which lead to strange and uninteresting questions. For example, the question "Why is the level of unemployment in Belgium 7.4 %, rather than the moon made of green cheese?" contrasts the topic with such a foil. Scientists will never ask this question. On the other hand, the question "Why is the level of unemployment in Belgium 7.4 %, rather than 10.7 % (which is the average level for the 27 EU member states)?" might very well receive attention. The green cheese problem points to the fact that the erotetic model not only allows for foils of the latter kind (the ones that lead to scientifically interesting questions) but also those of the former kind.

Second, the red herring problem. Like the green cheese problem, this problem is about excluding uninteresting elements from our explanations, only this time with respect to the relevance relation rather than the foil (Kitcher & Salmon 1987). There are countless relevance relations to choose from, but only a small portion of those single out answers we intuitively consider to be interesting explanations. If any relevance relation is allowed, then even if an adequate foil has been chosen, any true statement can explain any true contrast. Suppose the relevance relation is specified as "The answer must have more *r*s in it than the question". Then the question (to stick to our example): "Why is the level of unemployment in Belgium 7.4 %, rather than 10.7 %?" can be answered by the statement "Because red herrings have gills". Or (to borrow the original example of Kitcher & Salmon 1987) suppose we ask why John F. Kennedy died on November 22, 1963, rather than November 23. If there is no restriction on the relevance relation, then we cannot exclude statements based on astrological theory as answers: a true description of the positions of stars and planets at the time of Kennedy's

[34] This and the other figures are the unemployment rates for November 2012, as released by the European Commission on the website Eurostat. Retrieved 22-01-2013 from: http://epp.eurostat.ec.europa.eu/tgm/table.do?tab=table&language=en&pcode=teilm020&tableSelection=1&plugin=1

birth could count as a relevant answer for the question why Kennedy died on November 22, 1963 rather than on the next day. Clearly we want to exclude such inappropriate answers as scientific explanations (because scientists will never propose such answers). The red herring problem then, is how to exclude relevance relations that single out answers to why-question that, though true, are clearly not explanatorily interesting, because they are simply irrelevant.[35]

Although to my palate, the green cheese and red herring problems have a distinct armchair flavour, I believe that Kitcher and Salmon have raised a valid objection against van Fraassen's erotetic model. For Kitcher, pragmatics is not enough, and so he invokes the notion of significance as a relevance constraint (2001 p. 65). Of course, this new notion of significance begs for explication, and we shall take a closer look at it in part IV of this dissertation.

3.4 A different flavour of pluralism

So the upshot of Kitcher and Salmon's critique is that on van Fraassen's terms, any true proposition can explain any other true proposition. We will take a closer look at the green cheese and red herring problems, and some possible solutions to them, in part IV of this dissertation; nevertheless, we might concede at this point that a pluralist position is not very attractive from a philosophical point of view if there are no constraints *whatsoever* on what counts as an explanation. In this sense, van Fraassen commits himself to what we may term *anything goes pluralism* (to be defined below). The challenge then, is to make claims that are at once sufficiently general, or robust enough, to provide genuine insights into the explanatory practices of scientists, while avoiding the pitfalls of monism.

---

[35] It should be noted that some philosophers have construed incorrect examples of the red herring problem. Thus, it has been argued that if the relevance relation states that the answer must have three *r*s and two *g*s in it, then the answer "Red herrings have gills" explains why Bush is president (Khalifa 2004). This however is a wrong example, for 'having three *r*s and two *g*s' is a property of the explanans, *not* a relation between explanans and explanandum. The confusion stems, I think, from of a shift in terminology: Khalifa follows Risjord's substitution of the term 'relevance relation' with 'relevance criterion' (Risjord 2000a), and of course, though not a relation, 'having three *r*s and two *g*s' can be a criterion. Although this is not a serious problem (as I have just shown, the relevance criterion can easily by reformulated into "the answer must have more *r*s in it than the question"), nevertheless I feel that to avoid such confusions, it is best to stick to van Fraassen's original term; especially since the terminological shift was not made on principled grounds (Risjord 2000a, p. 71 note 4).

At this point, we find ourselves at the crossroads. If we adhere to van Fraassen's framework, it seems there are three possibilities open to us, as Kitcher informs us:

> Should we suppose there is a single set of genuine relevance relations that holds for all sciences and for all times? If not, if the set of genuine relevance relations is different from science to science and from epoch to epoch, should we try to find some underlying characterization that determines how the different sets are generated, or should we rest content with studying a particular science at a particular time and isolating the genuine relevance relations within this more restricted temporal and disciplinary area? (1989 p. 417)

The first of these three possibilities is represented by monistic approaches to scientific explanation.[36] As I have said in section 3.1, the lesson to be learned from our exposition of the various monistic theories devised in the twentieth century is that this is not a viable option.

The last possibility Kitcher mentions is what I have just labelled anything goes pluralism. On this account, it is the job of philosophy of science to simply describe what contrast classes and relevance relations happen to be used within a particular discipline at a particular time. We can delineate the different varieties of pluralism by providing yet more theses along the lines of the two we encountered in section 3.1. To refresh our memory, these were:

Minimal pluralism

A) There are no general exclusion rules with respect to scientific explanations.

B) There are no general preference rules with respect to scientific explanations.

[36] Of course, although the previous chapter also included a section on unificationism, clearly Kitcher thought his own project was not monistic in this sense, and in fact, the second option he presents in the quote above is the one he advocates. However, there are two observations to be made here. First, although I have drawn on Kitcher in describing unificationism (as he is considered the philosopher who has contributed most to developing that theory), one may of course choose to treat a philosophical position independently from its protagonists. In this case, one may simply summarize the monistic character of unificationism as the claim that *all* explanation is aimed at unification. Second, Kitcher *is* monistic in that he thinks that all explanations are deductive arguments (1989 p. 448).

Again, A and B are the minimal commitments to which explanatory pluralists ascribe, and to deny any one of them is to revert back to monism, either to weak monism (accepting A but rejecting B[37]), or to strong monism (rejecting both).

On this account, van Fraassen is certainly a pluralist, as he accepts both A and B. However, he seems to go beyond minimal pluralism, in that he does not place any restrictions on the relevance relation or the type of why-question that is to count as relevant. That is, besides A and B, he also subscribes to the more radical claims[38]:

Anything goes pluralism

C) There are no local exclusion rules with respect to scientific explanations.

D) There are no local preference rules with respect to scientific explanations.

Although I think that without any further restrictions, A and B make for a position that seems unambitious (or at least from a philosophical point of view unsatisfying), at least I believe that A and B are true – not so with C and D. A and B need to be supplemented with additional restrictions to make for a philosophically interesting position. In fact, they need to be supplemented with just the type of restrictions C and D deny are to be had. This, I take it, is the gist of the second option Kitcher presents in the quote above: to allow the relevance relations and contrast classes to vary from epoch to epoch, and from discipline to discipline, yet still be able to provide some local constraints. Kitcher himself tries to avoid anything goes pluralism by explicating some underlying characterization that determines how the different sets of relevance relations are generated, but of course one need not follow him on that specific route. In any case, as van Fraassen's particular brand of pluralism is not sufficient, we need to extend his account. I submit that we need to arrive at:

Robust pluralism

A) There are no general exclusion rules with respect to scientific explanations.

---

[37] An example of such a weakly monist position would be to grant that one cannot rule out both CL and CM explanations, but to insist that CL is always superior to CM. Note also that the reverse (rejecting A and accepting B) is also conceptually possible, for example if one would claim that there are exclusion rules to the effect that only CL and CM constitute genuine explanations, but that there are no general rules deciding which of these two is superior.

[38] Again, as far as I am aware of, van Fraassen never introduced additional restrictions to the model of explanation he presented in his 1980.

B) There are no general preference rules with respect to scientific explanations.

E) There are local exclusion rules with respect to scientific explanations.

F) There are local preference rules with respect to scientific explanations.

Of course, the main philosophical burden placed on anyone defending robust pluralism is to explicate the local rules referred to in E and F, and the robustness of robust pluralism will depend on how local the local exclusion and preference rules are. This will be discussed in chapter 14. First however, we need to be clear on the relation between robust pluralism, as a philosophical position on scientific explanation, and the pragmatic approach, as a meta-philosophical way of studying explanation (see note 5). This relation will be discussed in the next section.

Chapter 4
The pragmatic approach to explanation

4.1 Motivation and aims of the pragmatic approach

Now that we have familiarized ourselves with the most important theories about scientific explanation, including van Fraassen's pragmatic account, it is possible to clearly state the motivation for the pragmatic approach. Obviously, this motivation has to do with the controversy between monism and pluralism I have referred to in the previous chapter. Our starting point is the idea that scientific explanation is too diverse to be accounted for in terms of one overarching theory – in this sense, the pragmatic approach sides squarely with the pluralists. On the other hand, we have seen that there is a danger of lapsing into anything goes pluralism, which has the ultimate consequence of trivializing the debate about explanation. This is exactly where van Fraassen goes wrong.

This constitutes one of the main goals of the pragmatic approach: to say something informative about the explanatory practices of scientists, *without* claiming that all instances of explanation have some (formal) property in common. The erotetic model can serve as a starting point, but it is not enough if we want to say something positive and non-trivial about scientific explanations. To achieve this, in what follows we will take into account the epistemic interests or motives researchers have to pose certain questions. By doing so, one gains insights into certain features of explanations that, although they do not apply to every single instance of explanation (they remain local or domain-specific), nevertheless cover enough cases to significantly advance our understanding of it. In this sense, the pragmatic approach tries to arrive at a version of robust pluralism, as defined in section 3.4, effectively sailing a middle course between the Scylla of monism and the Charybdis of triviality.

Before moving on however, let me make one brief observation. One might say that the pragmatic approach is called an *approach* because it is situated at the meta-level – it

is not another *theory* of explanation, like CL, CM, U etc., but rather a *way of studying* explanation that employs such theories as tools. In effect, it is a meta-philosophical position. Nevertheless, the ultimate goal of the pragmatic approach, at least as it is used in the context of this dissertation, is to arrive at a view of explanation that conforms to the four claims of robust pluralism (A, B, E & F). That is, one's meta-philosophical commitments may have consequences for the theory one develops at the object level – and in fact, it is precisely the stalemate on the object level (the monism-pluralism controversy described in the previous chapter) that serves as a motivation for reflection at the meta-level. For now, let us contend that as far as the object-level is concerned, we employ minimal pluralism as a kind of working hypothesis, with the ultimate aim of developing it into robust pluralism in part V.

4.2 The pragmatic approach clarified: two guiding principles

The pragmatic approach to scientific explanation, as it has been developed at Ghent University over the past years, commits itself to two guiding principles.[39] The first principle is to argue for *domain-specific descriptive claims* about scientific explanation. Often, this involves looking closely at how scientists construct and test their explanations, and then offering a description of the explanatory practices in a given field or research tradition. Alternatively, one may select one subtype of explanations (such as causal or DN explanations), and offer an analysis of the structure and explanatory power of that subtype as it is constructed and used in a specific scientific field. Whatever the preferred method however, the arguments one can offer for domain-specific descriptive claims are empirical: they are based on case studies drawn from the scientific literature. The majority of the research that I will present in the second and third parts of this dissertation follows this first guiding principle. The advantage of this bottom-up approach is that the results are, in principle, accessible to the scientists themselves, and can thus potentially influence scientific practice.

This last point is important when considering the second principle, which is to argue for *domain specific normative* claims: by reflecting on the explanatory practices of scientists along the lines of the first principle, one can validate, defend or criticize certain explanatory strategies. As the aim of these normative claims is to improve explanatory practice, the accessibility of the bottom-up approach to scientists I referred

[39] These principles are adapted from chapter 2 of Weber, Van Bouwel & De Vreese (In press)

to in section 1.4 is a great advantage when arguing for context-dependant normative claims. Of course, it is important to take into account the epistemic interests behind explanations when arguing for these normative claims. These epistemic interests serve a dual purpose: they act not only as motives for scientists to search for and construct explanations, but also as motives for other parties (e.g. policy makers or the general public) to be interested in these explanations, and even provide funds for scientists to pursue them. In section 4.3, I will consider four of these epistemic interests (I will introduce others in chapters 12 and 13): *understanding* (the basic desire to increase one's knowledge), *improvement* (the desire to make things better), *prevention* (closely related to improvement, but more future-oriented, this is the interest to avoid undesirable situations), and *responsibility* (the desire to use an explanation to attribute moral and/or legal responsibility). Often, responsibility plays a role when we want to explain actions (cf. Weber & VanderBeeken 2005) or accidents (see the example below).

As we will see, it is the epistemic interests that determine which explanatory properties (e.g. depth versus breadth, predictive power, simplicity etc.) are important in a given context[40], and as such, they can serve as the evaluative criteria by which one judges the merits of particular explanations. For this reason, epistemic interests are a good tool to argue for normative claims. Below, I will consider an example to illustrate how this works.

4.3 An illustration: Richard Feynman on the Challenger disaster[41]

4.3.1 The disaster

Now that we have a grasp of the basic motivation and methodology behind the pragmatic approach, let us consider an example to illustrate how it works in practice. The example I shall consider is Feynman's analysis of the Challenger disaster.

[40] The relation between epistemic interests, types of explanation and the value of properties of explanations is investigated in Weber, VanBouwel & Vanderbeeken (2005) and Weber & Van Bouwel (2007), and will be considered throughout parts II and III of this dissertation.

[41] The following section (with the exception of subsection 4.3.4) is largely based on chapter 4 of Weber, Van Bouwel & De Vreese (in press).

On January 28, 1986, the space shuttle Challenger exploded, a mere 73 seconds after being launched from Kennedy Space Center in Florida, killing all seven members of its crew. The mission (51-L) was the 25th flight of the American Space Shuttle Program, and it was to have been a standard mission in several important respects. First, the Challenger was not the first space shuttle launched in the program (Columbia was the first), second, it had already served on nine successful missions, and finally, it carried a standard cargo (a satellite and other scientific equipment). In other respects, though, the mission was exceptional. First, January 28, 1986 was an unusually cold day, and second, on a more social note, the crew included Christa McAuliffe, who was supposed to be the first teacher in space – this last fact caused unusually extensive media coverage of the launch, and therefore also of the disaster itself.

After the disaster, at the behest of President Reagan, a presidential commission was formed to investigate the causes of the disaster. The commission was chaired by former secretary of state William P. Rogers. Among the other members of the commission were the retired astronaut Neil Armstrong (as vice chairman) and the eminent and Nobel Prize winning physicist Richard Feynman. In his book *What Do You Care What Other People Think* (1988), Feynman offers a personal account of the investigation. According to Feynman, the commission settled on two leading questions. The first question was: What physically caused the Challenger to explode? The second was: What went wrong in NASA that made the explosion possible? While both questions are motivated by the interest of understanding (they are aimed at increasing our knowledge), it is obvious that in the case of hand, understanding is a subsidiary goal: ultimately, the questions are motivated by prevention and responsibility.

4.3.2 A CM explanation

Of course, to answer the first question, it was necessary to obtain some technical data. A lot of research went into the workings of three crucial systems: the solid rocket boosters (SRBs), the space shuttle main engines (SSMEs), and the flight electronics. Early on in the investigation, it transpired that there was something seriously amiss with the so-called O-rings; rubber rings that were supposed to seal the aft field joint of the right SRB. The SRBs are comprised of segments connected to each other with a Tang and Clevis joint: each segment has a U-like shape at the top, allowing the bottom of the segment above it to slide neatly in place. The two rubber O-rings that seal these joints are designed to be very flexible, so as to prevent gas leaking through from the SRB. It turned out however that it was unclear just how the rubber seals would react to the cold temperature on the day of the launch. Moreover, some technicians had already expressed doubts about the quality of the O-rings before the launch.

Philosophically speaking, Feynman's answer to the first question was obtained by searching for a CM explanation, tracing the causal pathways leading up to the effect – the explosion. This involved questioning technicians about the behaviour of the relevant parts of the space shuttle's rocket boosters and main engines under normal circumstances, and under the specific circumstances obtaining on the morning of the launch, namely under extreme cold. The explanation detailed how gas could leak from the SRB trough the joints, causing fire and an explosion, as Journalist Nick Greene sums up:

> The commission's report cited the cause of the disaster as the failure of an "O-ring" seal in the solid-fuel rocket on the Space Shuttle Challenger's right side. The faulty design of the seal coupled with the unusually cold weather, let hot gases to leak [sic] through the joint. Booster rocket flames were able to pass through the failed seal enlarging the small hole. These flames then burned through the Space Shuttle Challenger's external fuel tank and through one of the supports that attached the booster to the side of the tank. That booster broke loose and collided with the tank, piercing the tank's side. Liquid hydrogen and liquid oxygen fuels from the tank and booster mixed and ignited, causing the Space Shuttle Challenger to tear apart. (Greene n.d.)

4.3.3 Some contrastive questions

Although the CM explanation, as summed up by Greene, gives us a satisfactory answer to the first question the commission posed, it does not answer the second question, which is about the internal workings of the NASA. It did however allow the commission to formulate two further questions:

1) Why was the Space Shuttle Program continued, rather than put on hold until the problems of gas leakage and erosion were solved?
2) Why was the launch of the Challenger not postponed to a warmer day?

Although related to the original second question (What went wrong in NASA that made the explosion possible?), these questions are more specific and presuppose that we already have an answer to the first question. It is worthwhile at this point to note a property that both these questions have in common: they are contrastive, as described in section 3.2. In the case at hand, both questions compare a state of affairs that is the case, with a state of affairs that is not the case. In effect, they tell us how NASA really functioned, and contrast this with how NASA might have functioned, and in doing so,

they tell us how things should have been different in order to obtain the alternative state of affairs.

Moreover, with an eye to what is still to come in this dissertation, note that we are not entirely neutral with respect to this partitioning in terms of truth: with hindsight, we would rather have that the topics were false and the foils true (that is, we would rather have that the Challenger disaster did not happen). Here we see something emerging that will become important later on in this dissertation, namely that we can have certain attitudes when it comes to contrastive questions: we might want, desire, hope, or expect certain propositions to be true, and others to be untrue. In turn, these attitudes can act as motivations, or interests, to ask certain questions. In the case of the Challenger, the contrastive explanation-seeking questions are obviously motivated by the interest of prevention, and the hope is that answering them will make future Shuttle launches safer. Again, in part IV, the relations between interests and explanation-seeking questions will be addressed in more detail.

Moving on, let us focus on the first follow-up question. Further into the investigation, it was found that there had been gas leakages in some of the seals during a number of previous missions, resulting in clearly visible black marks, and so-called 'erosion' spots where the O-rings were partially burnt. This evidence indicates that the O-rings did not satisfactorily perform the function (namely to seal the joints) they were designed for (the "faulty design" referred to by Greene in the quote above). According to Feynman, the way NASA reacted to these incidents was wrong, even going so far as to compare their risk analysis methods and risk management strategies to Russian roulette:

> The phenomenon of accepting for flight, seals that had shown erosion and blow-by in previous flights, is very clear. The Challenger flight is an excellent example. There are several references to flights that had gone before. The acceptance and success of these flights is taken as evidence of safety. *But erosion and blow-by are not what the design expected. They are warnings that something is wrong*. The equipment is not operating as expected, and therefore there is a danger that it can operate with even wider deviations in this unexpected and not thoroughly understood way. The fact that this danger did not lead to a catastrophe before is no guarantee that it will not the next time, unless it is completely understood. When playing Russian roulette the fact that the first shot got off safely is little comfort for the next. The origin and consequences of the erosion and blow-by were not understood. They did not occur equally on all flights and all joints; sometimes more, and sometimes less. Why not sometime, when whatever conditions determined it were right, still more leading to catastrophe [sic]?
>
> In spite of these variations from case to case, officials behaved as if they understood it, giving apparently logical arguments to each other often depending on the "success" of previous flights. For example, in determining if flight 51-L was safe to fly in the face of ring erosion in flight 51-C, it was noted that the erosion

> depth was only one-third of the radius. It had been noted in an experiment cutting the ring that cutting it as deep as one radius was necessary before the ring failed. Instead of being very concerned that variations of poorly understood conditions might reasonably create a deeper erosion this time, it was asserted, there was "a safety factor of three." This is a strange use of the engineer's term "safety factor." If a bridge is built to withstand a certain load without the beams permanently deforming, cracking, or breaking, it may be designed for the materials used to actually stand up under three times the load. This "safety factor" is to allow for uncertain excesses of load, or unknown extra loads, or weaknesses in the material that might have unexpected flaws, etc. If now the expected load comes on to the new bridge and a crack appears in a beam, this is a failure of the design. There was no safety factor at all; even though the bridge did not actually collapse because the crack went only one-third of the way through the beam. The O-rings of the Solid Rocket Boosters were not designed to erode. *Erosion was a clue that something was wrong. Erosion was not something from which safety can be inferred.*
>
> There was no way, without full understanding, that one could have confidence that conditions the next time might not produce erosion three times more severe than the time before. Nevertheless, officials fooled themselves into thinking they had such understanding and confidence, in spite of the peculiar variations from case to case. (Feynman 1986; italics added)

This lengthy quote constitutes a neat summary of Feynman's answer to the second question. The sentences put in italics refer to those points in the assessment procedure where the evidence, according to Feynman, raises red flags that should have been noticed by, and influenced the decisions of, the NASA space agency. Feynman answers the first of the follow-up questions in terms of 'officials fooling themselves', that is, the officials should have seen the red flags raised by the gas leakages and the erosion spots and acted accordingly.

4.3.4 Evaluating Feynman's explanations

In analysing the case of the Challenger disaster, we see how the two guiding principles of the pragmatic approach work in practice. They involve having a close look at how a researcher constructs his explanations. In the case at hand, we look at how, in the interest of understanding, Feynman includes technical data from multiple sources and pieces them together in a story that describes the salient causal pathways leading up to the explosion. The domain-specific descriptive claim here is that Feynman constructs a CM explanation. As far as the CM explanation informs us about the physical causes of the disaster, and so serves the interest of understanding, we can also make the normative claim that it is a good explanation.

But there is another normative claim we can make regarding the case of the Challenger. As we have seen, the questions about the workings of NASA are motivated by the interest of *prevention*: we want to prevent a similar event from occurring in the future. This interest provides us with the means to normatively evaluate Feynman's explanations. As these explanations laid bare some important points in the decision making process where improved sensitivity of the decision-makers to the warning signals could have made a difference, the explanations can lead to recommendations for future policies, which indeed they did: the commission made a number of recommendations, and NASA wasted little time in reporting back to the president how and when they would implement these recommendations (for details, see Fletcher 1986). Thus, we see here how policy makers can be motivated by the interest of prevention to take an interest in the explanation offered by Feynman. To the extent that the contrastive explanation serves the interest of prevention, we can make the domain-specific normative claim Feynman's explanation, consisting of a combination of the CM explanation and a folk-psychological one describing the attitude within NASA, is a good explanation. Thus, although the CM explanation by itself makes for a good explanation, as it increases our understanding, it is also part of a wider explanation serving the interests of prevention, and, ultimately, the attribution of moral responsibility to NASA officials.

Thus, we see here how epistemic interests serve their dual function as motivations for providing and evaluating explanations. Finally, we can appreciate that the resulting picture is indeed pluralist: the pragmatic approach acknowledges the value of CM explanations as well as folk-psychological explanations citing the mental attitude of NASA officials. Thus, this example also illustrates the instrumentalism of the pragmatic approach: it does not put any a priori, general restrictions on what types of explanations exist, and what makes for a good explanation, instead allowing different types of explanations to co-exist and complement each other in a given case.

4.4 How to further develop the pragmatic approach

Hopefully, the previous sections have given the reader an impression on how the pragmatic approach to scientific explanation works in practice. I have explicated some of the explanation-seeking questions that were asked by the commission investigating the Challenger disaster case, and identified some of the epistemic interests that serve both as motivations for asking these questions, and as criteria by which to evaluate the resulting explanations. In this way, we can make both domain-specific descriptive and domain-specific normative claim about particular explanations found in the literature.

This last consideration brings to light an important point about the methodology of the pragmatic approach I already mentioned in section 1.4: *we develop and expand it as we go along*. By analysing a particular debate about explanation in a particular discipline, we adjust and refine the tools we have at our disposal, and if need be, introduce new ones. At the close of this chapter, let me stress the importance of the bottom-up methodology of the pragmatic approach. As we have seen in the case of the Challenger disaster, it is by considering a concrete example of scientific explanation that we learn what conceptual tools (question-types, epistemic interests, etc.) we need to introduce. Simultaneously, this methodology lends these conceptual tools their justification, ensuring that the pragmatic approach stays true to the naturalistic commitment I endorsed in section 1.1. It is with this in mind that I made the commitment in section 1.4 to illustrate controversial claims with examples taken from actual scientific practice rather than everyday examples. In a sense, the examples are themselves the arguments for the descriptive claims I make.

This example is a clear illustration of how the pragmatic approach has hitherto been used to analyse scientific explanations. As I have said in the introduction however, I have made some specific contributions to the approach: besides further developing the theoretical side of the approach, in particular the relation between different question-types and the relevance of the answers to these questions (see part IV), I have also applied to approach to the relation between CL explanations and mechanistic explanations, which is the subject of the next part, and to the issue of model explanations, which is the subject of part III.

Part 2

Covering law versus mechanistic explanations in the cognitive sciences and biology

Chapter 5
Introducing cognitive science: functional and mechanistic explanations

5.1 Delineating the field: two outstanding features of cognitive science

The term 'cognitive science' as it is used today is really an umbrella term. It groups together a wide range of disciplines that, to a greater or lesser extent, all have cognition[42] as their subject. These include: psychology, linguistics, computer science, and neuroscience, to name a few. In turn, each of these disciplines themselves can be subdivided into different branches. For example, psychology branches out into disciplines like behavioural, cognitive and social psychology; computer science covers information theory, programming and artificial intelligence, among others; and neuroscience includes the localization of cognitive capacities in brain structures. Also, there has been considerable input from philosophers – especially philosophers of mind, language, science, as well as from epistemologists.

Thus, cognition is studied by many different disciplines. Moreover, rather than operating in isolation from each other, these disciplines interact and exert mutual influence on each other.[43] One of the most striking features of cognitive science then is

[42] Where the term 'cognition' is also highly heterogeneous, see section 5.2.

[43] Of course, this goes against the idea, once popular among functionalist philosophers of mind, of methodological dualism: the claim that psychology and neuroscience operate as entirely distinct and autonomous disciplines, each with their own research agendas, explanatory strategies and explananda (Fodor 1974, 1997). Surveying the field, interaction between the different disciplines comprising what we call

its highly interdisciplinary character. Another is its *dynamic progress*. Over the course of its relatively short history, it has seen three periods of rapid development that are commonly referred to as *revolutionary* by commentators.[44] The first of these is often called the *cognitive revolution* (Gardner 1985). During this period, the field of artificial intelligence was founded by theorists such as John McCarthy, Allen Newell and Herbert Simon, while simultaneously, due in no small part to the work of Noam Chomsky, experimental psychology saw the final overthrow of the behaviourist research programme. The second 'revolution', occurring during the 1970s and 1980s, was constituted by the introduction of what one might call the brain-based view. This brain-based view was inspired by research on new computational architectures, resulting in what we now know as connectionist networks, coupled with the development of new imaging techniques in neuroscience. Finally, the past twenty years have seen the rise of a movement that could be labelled *dynamicism*. Its creed is neatly summed up by Wheeler: "...cognitive science needs to put cognition back in the brain, the brain back in the body, and the body back into the world" (2005 p. 11). In other words, dynamicism eschews the practice of traditional cognitive science to analyse cognition in abstraction from its neural and bodily surroundings. To view cognition as a constant, dynamic and interactive process between mind, brain and body is the upshot of so-called *embodied cognition*. Scientifically speaking, the dynamicist movement was inspired by the introduction of new mathematical models of cognition (dynamical cognitive science) to compete with the old computationalist (both classic AI and connectionist) models. These dynamical models use differential equations to simulate cognitive capacities, – as we will see in the next chapter, this has explanatory implications. Philosophically speaking, the movement has extended the boundaries of cognition beyond the brain and body, and into the world (Wilson 2004; Clark 1997, 2008; Clark & Chalmers, 1998).[45]

The details of these two features of cognitive science need not detain us here. The interdisciplinary nature of cognitive science has been commented on many times in the

---

cognitive science seems a given, whether or not this suits our own particular philosophical perspective (cf. Gervais 2012a).

[44] It has been noted though that the term 'scientific revolution' has suffered from severe inflation in post-Kuhnian philosophy of science: "...the term (...) has been used with increasing frequency in discussions of scientific change, and the magnitude required of an innovation before someone or other is tempted to call it a revolution has diminished alarmingly" (Ramsey, Stich & Garon 1990 p. 499). With this in mind, I will not take a stance here on whether or not the three periods in the history of cognitive science referred to *are* revolutions in a philosophically interesting (Kuhnian or otherwise) sense; I simply record the fact that they were tumultuous periods of research during which numerous breakthroughs were made, and that they have consequently been termed revolutions by commentators.

[45] Besides abolishing the isolated Cartesian subject, dynamicism has also attempted to break the deeply entrenched philosophical habit to view cognition as a kind of *representation* – more on that in section 6.1.

literature, and its tumultuous history has likewise been chronicled extensively. In the context of the present dissertation, what matters is that when it comes to explanation, these features imply that the explanations furnished by cognitive scientists are often very heterogeneous. They may draw on resources from different disciplines, they may approach the same explanandum from different angles, and they interrelate in different ways (e.g. competition, complementation etc.).

5.2 Subject matter and explananda

Like so many other disciplines, cognitive science gets its name from its subject matter. Yet what does cognition mean? It has been noted that the rise of the term 'cognition' among psychologists dates back to the mid-1950s, that is, to the demise of behaviourism. Although it was clear that behaviourism as a research programme for psychology could not succeed and that behavioural data could only be understood in light of the underlying mental processes, many experimental psychologists were still reluctant to employ overly mentalistic terms, and so began to use the term cognitive instead (Miller 2003). However, over the years the bias against mentalistic vocabulary waned, and it is now again acceptable to talk about memory instead of repetitive learning, or speech production instead of verbal behaviour, etc.

As mentioned in the previous section, the counterrevolution of psychologists against behaviourism was accompanied by developments in artificial intelligence and computer science, most notably in the area of computational modelling, i.e. the efforts of programmers to simulate mental processes using computers. The tendency was to interpret many cognitive achievements as a kind of *information processing*. As is widely known, classic computationalism with its serial, brittle architecture did not enjoy its dominant status for long: connectionist networks, and later, dynamic systems theory, challenged prevailing ideas about the architecture of the mind. Moreover, as the neurosciences progressed and new techniques to map cognitive capacities onto neural substrates (fMRI, PET etc.) developed, the view became increasingly widespread that, *pace* the functionalist consensus of the 1970s and 1980s, implementational details *do* matter.

Yet despite all the intellectual turmoil the cognitive sciences have gone through in the past decades, some things have remained the same. What has remained are the *explananda* of the cognitive sciences. Memory, arithmetic skills, pattern recognition, spatial recognition, speech acquirement and production, but also more specific applications of these abilities, such as the ability of rats to navigate through a Morris water maze or of hawks to spot a rabbit on the ground – all these skills, as they are

performed by adults, children or animals, are among the things that cognitive scientists try to explain. In short, *capacities* are an important type of explanandum[46].

As should be evident from the preceding paragraph, explananda in the cognitive sciences can be *particular facts* as well as *regularities*. We might be interested to know why some particular individual possesses a certain capacity, or we might want to know why a group of individuals share the same capacity, or we might be interested in the differences between individuals, for example when some members of a set do not possess a capacity that other members do have. Likewise, in game theory we might want to explain an individual's choices in a particular test run of the ultimatum game, or we might want to explain the performance of a specific group – say, people over fifty.[47]

5.3 Models: functional and mechanistic explanations[48]

As capacities are such an important type of explanandum, it is worthwhile to take a moment to reflect on the explanatory strategies employed by cognitive scientists to explain them. Of particular importance here is that capacities can be the subject of different types of explanation-seeking questions. We might be interested in the reasons why an individual or group exhibits a certain kind of capacity. For example we might ask: "Why do humans see depth?" The answer here might, for example, be couched in evolutionary terms, and explain the occurrence of the capacity to see depth in terms of the advantage this particular trait gives to the species.

However, we might also be interested in the way a capacity is realized. That is, rather than a why-question, we might actually ask: "How do humans see depth?"[49] Here, we

[46] Indeed, Cummins claims that in psychology, capacities constitute the "...primary explananda..."; and he goes on to say that "Understanding (...) capacities is what motivates psychological inquiry in the first place" (2000, p. 121). I don't know whether this is true (how does one rank explananda?), but I concur they are important, if for nothing else, because they are ubiquitous.

[47] As Cummins (2000) points out, in psychology these regularities are often referred to as *effects*. The recency effect, the primacy effect, the McGurk effect and the framing effect are just some examples.

[48] Parts of this section will appear in Gervais (in print), Gervais & Looren de Jong (in print), and have appeared Gervais & Weber 2013a. As I have said in section 1.5, although functional explanations are not of particular importance to this part (I will need them primarily in part III), I thought it most natural to introduce them here, in opposition to mechanistic explanations – the two types of explanation are regularly contrasted in the literature (e.g. Craver 2001).

want some description of how the capacity is brought about. This description can take the form of a *functional model*.

Functional models work by means of decomposition. They are used to explain a capacity by breaking it down into sub-capacities or -functions, and then showing how the overall capacity is a result of the organization of these sub-functions (Cummins 1975). A useful metaphor here is that of the assembly line. Consider a factory churning out radios. This factory effectively performs the function of taking parts as input and producing radios as output. This function can be explained by dividing the assembly process into several sub-routines carried out by workers standing alongside a conveyor belt, where each subsequent worker adds a specific component to the radio, until the finished product appears at the end of the belt, ready for transport. Once we know all the sub-routines that make up the assembly process, and understand the way they are organized (the order in which the parts are added) we can explain how the factory performs its function by means of a flow-chart or box diagram.

This explanatory strategy was widely used in the cognitive sciences, especially in the 1980s and 1990s. Cognitive capacities like memory storage, face recognition and numerical cognition were explained by construing models of how these capacities might be divided up into sub-functions. In psycholinguistics for example, a particularly influential functional model for the capacity of speech production was offered by Levelt (1989). Roughly, the process was divided into three steps: first, the person conceptualizes what he wants to say, second, he formulates this into language (this step is in turn divided into two sub-tasks, one of lexicalization, which produces the words needed, and one of syntactic planning, which provides order and grammatical structuring to these words) and finally, he engages in articulation (see figure 1).

[49] Thus, how-questions, like why-questions, can be scientifically legitimate; hence I reject van Fraassen's first claim as listed at the end of section 3.2. As the erotetic model of explanation only covers why-questions, we need to consider this issue in more detail. I will do so in chapter 13.

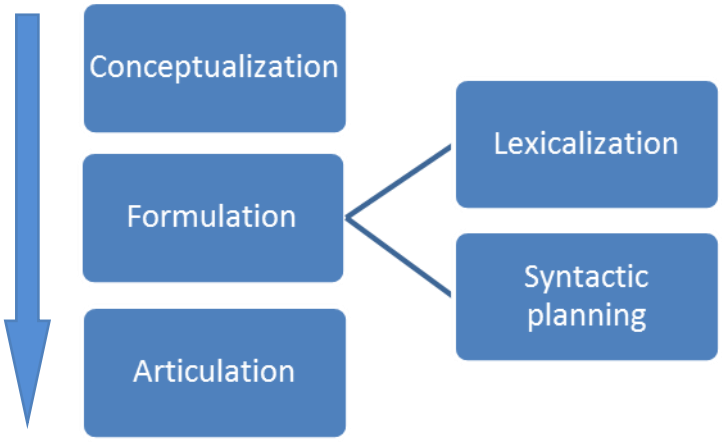

Figure 1 An adaptation of Levelt's functional model of speech production. The sequence of operations is depicted vertically, with one step (formulation) being divided into two sub-routines (lexicalization and syntactic planning).

Of course, this is only a rough sketch of how the capacity might be realized, but it need not be wholly speculative. For example, the distinction between lexicalization and syntactic planning may be grounded in experimental evidence: some test subjects might be able to produce the right words, yet fail to put them in the correct order. In general, functional models need not be merely phenomenal (input-output mapping devices): with respect to the partitioning of a capacity into sub-routines, one can be detailed or abstract, and this partitioning might be supported by experimental evidence to a greater or lesser degree (I will consider this point in detail in chapter 10).

Given the centrality of capacities as explananda, and the popularity functional models enjoyed among cognitive scientists, philosophers have made various attempts to provide an analytic account of the notion of capacity or function[50,51] – one particularly influential attempt was Robert Cummins' so-called 'causal role' (henceforth CR) account of functions. Consider Cummins' definition (1975 p. 762):

CR: X functions as a ϕ in S (or the function of X in S is to ϕ) relative to an analytic account A of S's capacity to ψ just in case X is capable of ϕ-ing in S and A appropriately and adequately accounts for S's capacity to ψ by, in part, appealing to the capacity of X to ϕ in S.

The core idea of Cummins is this: a capacity is analysed in terms of two or more distinct and simpler sub-capacities, which in turn can be analysed in terms of still simpler sub-

[50] Among these are dispositional (Bigelow & Pargetter 1987) and etiological accounts (Mitchell, 2003).

[51] Although these authors (and Cummins) talk about an account of 'function', due to the heavy metaphysical implications of that notion (especially in the context of biology) I shall mostly stick to using the term capacity.

capacities. Thus we explain a capacity by detailing how certain sub-capacities in a certain organization contribute to the overall capacity exhibited by the system, that is, to the original explanandum.

Outside of cognitive science, this explanatory strategy was greeted with considerable enthusiasm by the philosophical community, who saw in it a naturalistic way to account for higher-level capacities by letting the process of division bottom out at a level where the sub-capacities are so simple they are no longer considered problematic, thus dispelling any worries about homunculi (Cummins 1980; Dennett 1987).

Yet however much informed a functional model might be, there is one issue with respect to which it remains silent: it does not include any information about what actually performs all these subtasks. To put the point differently, it specifies functions, but not the realizers of these functions. In the example of the assembly line, imagine that in another factory, the different assembly tasks are realized by robots instead of workers. From a certain level of abstraction, the two factories are functionally equivalent, as they both perform the function of taking in parts as input and producing radios as output.

Special emphasis can be placed on those functions that are 'the same' across many different implementations, the idea being that at a certain level of abstraction, two systems that are physically different can nevertheless be functionally equivalent (the so-called multiple realizability of functions). In cognitive psychology and the philosophy of mind, functionalism is a thesis about the nature of mental processes and their realization in physical processes. Functionalism sought to combine the metaphysical position of materialism or physicalism with the autonomy of the special sciences (Fodor, 1974). The combination of these two claims amounts to the position of non-reductive physicalism, which is probably the dominant position in philosophy of mind today.

To summarize the main conceptual point, a functional model explains a capacity by breaking it down into smaller sub-capacities, and then showing how those sub-capacities are organized to produce the original capacity. More formally, if we want to explain a capacity *C* of a system *S*, we have to construct a functional model *M* (account A in Cummins' definition) which performs *C*, such that for each input, output and input-output relation in *S* there is a corresponding input, output and input-output relation in *M*.

However, what was once hailed as an advantage, namely the absence of information about the realizers of a (sub-)function, is now increasingly criticised as a weakness. To be sure, functional models may succeed in correctly mapping the input-output relation of the target system, and for the purposes of control or prediction this may suffice, but does that make the model explanatory? Even though a particular partitioning of a function into sub-routines is supported by evidence, if we want to understand how *we*, as humans, perform some kind of cognitive capacity, it seems imperative that we know

something of the brain regions involved. Too often, the critics say, researchers are at a loss about what is really behind the boxes in their diagrams. For heuristic purposes, e.g. when we are just mapping out a certain capacity, this may be fine,[52] but if the original status of these boxes as mere placeholders is forgotten, they only serve to mask gaps in our understanding – hence the derogatory term 'boxology' that is sometimes applied to pure functional analysis (Craver 2001, 2006).

Following Machamer et al.'s seminal paper *Thinking about Mechanisms* (2000), a growing body of literature is devoted to an alternative approach to explaining cognitive capacities: *mechanistic explanations* (Glennan 2002, Bechtel & Abrahamsen 2005, Craver 2007). Although the notion *mechanism* is inherited from early modern philosophy, its meaning has departed from the traditional understanding of mechanisms: rather than physical systems involving gears, flywheels, springs or colliding particles, they are viewed as conceptual collections of entities and activities that are organized such as to produce a certain phenomenon, or, in our terms, capacity. Researchers aim to *explain* a capacity by describing the responsible mechanism. In other words, explaining a capacity consists of providing a model of the mechanism responsible for that capacity. Examples of disciplines in the life sciences where this explanatory strategy is used include (among others) neuroscience (Craver 2007; Bechtel 2008), cell biology (Bechtel 2006), genetics (Darden 2006), molecular biology (Darden & Tabery 2009), chemistry (Ramsey 2008) and psychology (Bechtel 2007).

Like functional explanations, mechanistic explanations decompose the target capacity into several sub-capacities. Unlike functional explanations however, mechanistic explanations also incorporate information about *what* performs a certain (sub-)function. They explain a capacity of a system by modelling the mechanism responsible for it: its entities, activities, and the way these entities and activities are organized all come into play.[53] Accordingly, Machamer et al. famously define a mechanism as "...entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (2000, p. 3). Of course, the model describing the mechanism need not be complete. Complete descriptions only serve as a regulative ideal: the degree of completeness required depends on our purposes at the time.

---

[52] See Machamer et al., who write that a mechanistic explanation typically starts by providing a mechanism sketch, which is "...an abstraction for which bottom out entities and activities cannot (yet) be supplied or which contains gaps in its stages. The productive continuity from one stage to the next has missing pieces, black boxes, which we do not yet know how to fill in" (Machamer et al. 2000 p. 18).

[53] Another way to put the difference is that mechanistic explanations, besides decomposition, also in involve localization, where the latter notion is understood as the identification of activities with parts (Bechtel and Richardson 1993).

So how do mechanistic explanations work? To explain a systemic activity of a system (e.g. a cognitive capacity in humans), one identifies the mechanism responsible for it. One level below that of the overall activity of a mechanism, we find its components or working parts, where components refers to activities as well as entities (Darden 2006). These parts may themselves be mechanisms, whose parts are on a still lower level. Here levels can only be assigned locally, with respect to a particular mechanism: they are developed for specific exemplars and there is no way or need to anchor them in a universal, basic level of nature (Bechtel & Abrahamsen, 2005). One of the reasons for this is that mechanistic explanations typically incorporate contextual and/or environmental data, which will vary from case to case. The point is that the mechanisms are realized via a constituency relation: the parts and their organization, on whatever level they might be, generate the mechanism.

Let us briefly consider the example of Stricker's and Verbalis's mechanistic explanation of fluid homeostasis as cited by Craver (2007 p. 9). After eating salty food, or after sweating without replenishing the lost water, the level of plasma osmolality rises. This causes vasopressin to be released, which in turn helps the body to conserve water in a number of ways, and evokes the feeling of thirst. In this example, an explanandum (osmoregulation) is explained by decomposing it into the different sub-routines and operations of different mechanisms (e.g. the pituitary releasing vasopressin). These mechanisms are located at different levels, from behavioural to molecular.

As this example shows, mechanistic explanations assume a very local account of levels, and are often highly inter-level. The mechanistic explanation of fluid homeostasis "...oscillates up and down in a hierarchy of mechanisms to focus on just the items that are relevant..." (Craver 2007 p. 10). It is important to note that there seems to be a determination relation between fluid homeostasis, which is an activity of the body, and several operations conducted by different parts of the body. These parts and their activities 'fix' the activity of fluid homeostasis of the body as a whole. As for the inter-level character of the explanation, because it cites parts and operations on different levels, it seems that once the explanandum is fixed, there are no principled restrictions on the relevant explanatory level possible any more (Bechtel 2009). Moreover, as the mention of parts already suggests, mechanisms are composed of hierarchically related entities and operations, and as such are thought to stand in a *constitutive* or *mereological* relation to their parts and activities.

Although, as we have seen, mechanistic explanations share the decomposition strategy with their functional predecessors, the insistence on information about the entities involved sets them apart. A Cummins-style analysis limits itself to specifying operations, yet, as Machamer et al.'s definition already makes plain, according to the mechanists, one must also include information about the parts that perform these operations. They speak of entities and activities, or parts and operations that have to be specified. Thus, although the mechanists explicitly build upon Cummins' CR account

(Craver 2001), they supplement it with the additional requirements. Bechtel & Abrahamsen formulate it as follows:

> A mechanism is a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena. (2005, p. 423)

Craver puts it as follows:

> [M]echanisms are entities and activities organized such that they exhibit the explanandum phenomenon. (2007, p. 6, italics removed).

Let me briefly summarize our terminology. First we have the following convention:

> A *capacity* is a system's ability to exhibit some kind of behaviour.

I assume (this is an ontological assumption, not a terminological convention) that such capacities are realized by *some* underlying mechanism, even though we may be ignorant with respect to its details. A mechanism is defined as follows:

> A *mechanism* is a collection of entities and activities that are organized such that they realize the capacity.

For the use of mechanisms in explanations (the 'analytic account A' in Cummins' 1975 definition), we have the following convention:

> A *mechanistic explanation* of a biological capacity is a description of the underlying mechanism.[54]

Thus, we see how the seemingly innocent observation that capacities are an important type of explanandum for cognitive scientists, leads to complex issues about models. If the pragmatic approach is to be applied to the explanatory practices of cognitive scientists, it is pivotal that the explanatory toolbox is extended in such a way as to make sense of these different kinds of model explanations.

[54] Thus, I employ what might be called an *epistemic* rather than an *ontic* conception of mechanistic explanations (cf. Wright 2012).

5.4 Explanatory inferences in cognitive science

As we will see in the next chapters, the CL model of explanation is no longer thought to be of great use to understand the explanatory practice of either cognitive science or biology these days, as philosophers tend to focus on mechanistic explanations[55]. However, there are at least two reasons to question this state of affairs. The first is the rise of dynamicism I referred to in section 5.1, particularly the mathematical models of dynamical cognitive science. The way these models describe and simulate certain patterns of behaviour in terms of differential equations, has led some researchers to argue that dynamical cognitive science in fact provides CL explanations (Chemero 2009, Chemero & Silberstein 2008, Walsmley 2008). In the next chapter, I will evaluate this claim.

But there is a second, more basic reason to question the dismissal of CL explanations in the cognitive sciences. This is that, at least at an intuitive level, a certain type of inferential explanations does occur, especially in the more traditional approaches to human cognition, (i.e. in pre-dynamicist approaches). A common tread recurring through most of these inferential explanations is that many cognitive capacities can be understood in terms of *mental representations* and *computational processes*, where these representations can be viewed as propositional attitudes, rules, images or concepts, and the computational processes as deduction, abduction, serial processing, or synaptic weight distribution, according to one's favourite theory of mind. Nevertheless, whatever one takes the nature of mental representations and computational processes to be, it is possible to construct inferential explanations that have certain formal properties in common.

For example, suppose we ask the generic question: "Why do people exhibit intelligent behaviour?" where this intelligence can be understood in terms of the cognitive capacities we have considered so far (e.g. speech production). In connectionist terms, the explanatory inference might look something like this:

(1) People have a connectionist network, i.e. processing units that have synaptic (excitatory and inhibitory) connections to each other

(2) People have procedures for spreading activation across these networks

(3) People have procedures for adjusting the weights of the synaptic connections

[55] See section 8.1. I will consider (and indeed, take issue with) the current appreciation of the CL model when applied to cognitive science and biology in more detail in chapters 6 to 8.

(4) Together, the procedures for spreading activity and adjusting the weight of the synaptic connections produce intelligent behaviour

while on classic, rule guided computationalist terms, the explanatory inference might look like this:

(1) People follow mental (conditional) rules

(2) People procedures to follow these rules when searching for solutions

(3) People have procedures for generating new rules

(4) Together, the procedures for following and generating rules produce intelligent behaviour

Hence, people exhibit intelligent behaviour. This may sound highly abstract, yet it does not stray far from how certain philosophers and cognitive scientists explained all sorts of intelligent behaviour in the past. For example, a rule-based view on language acquisition may lead to exactly this sort of explanatory pattern, where language acquisition consists of mastering the appropriate rules, from basic ones governing the semantics of everyday words to more complex syntactical rules, (e.g. rules governing the formation of compound sentences). Of course, many of these theories of mind and their concepts are outdated (at present, few theorists still believe that the mind operates as a computer with a serial architecture, for example) and in fact the above explanations may be wrong.

However, there is an important lesson here. What these explanations have in common is that one's cognitive capacities are explained by reference to generalized statements about mental representations people are supposed to have, and the computational processes acting on these representations. Thus, although not all explanations in cognitive sciences are arguments, in line with the pluralism developed in the previous chapter, we would be ill-advised to exclude such explanatory inferences. It seems that besides model explanations, Hempelian-like CL explanations also ought to be accommodated if the pragmatic approach is to make sense of the explanatory practices of cognitive scientists.

Chapter 6 Dynamical cognitive science

6.1 Dynamical cognitive science

In this chapter, I will argue that some explanations in dynamical cognitive science are a *specific type* of CL-explanations. I emphasize the phrase 'specific type' to indicate that these explanations do not strictly conform to Hempel's rules. Rather, as we will see, they are non-deductive causal CL explanations using a default rule, instead of strict deductive arguments using exceptionless laws of nature. I will reach this conclusion in dialectical opposition to a recent critical evaluation of explanation in dynamical cognitive science by Walmsley (2008). According to Walmsley, these explanations strictly conform to Hempel's CL model, and as such, he suggests they are problematic in some respects – a conclusion I reject. Thus, evaluating Walmsley's analysis is not a goal in itself. Rather, the main goals of this chapter are i) to give an accurate description of the explanatory practices in dynamical cognitive science, and ii) to evaluate the usefulness of these explanations. However, one cannot achieve these goals without having some prior knowledge of dynamical cognitive science itself.

Dynamical cognitive science is an application of *dynamic systems theory*, which studies and describes the behaviour of complex dynamic systems with the help of mathematical tools such as differential and difference equations. It also involves elements from computer science, most notably computer models and simulations, which can be used to describe all kinds of dynamic systems. Some of these systems can exhibit a property known as *complexity*, meaning they are not simply a collection of static entities that can be studied individually, but rather dynamic networks of entities and interactions between these entities. Dynamic systems theory then tries to describe how relationships between parts give rise to the collective behaviour of (both natural and artificial) systems. Its applications are many, and it can be used to study systems found in diverse fields, such as artificial intelligence, economics, chemistry, geology and social psychology. In these fields, complex systems are described in terms of a space of possible states, the so-called *state space*, and particular trajectories through that space,

where these trajectories can in turn be described using equations. This mathematical character, together with the possibility to run computer simulations, makes dynamic systems theory particularly apt for the purpose of prediction, and is therefore often used to analyse systems that are of particular interest to human society. For example, in the case of the last two disciplines mentioned, geology and social psychology, dynamic systems theory is used to attempt to predict earthquakes and the behaviour of crowds in confined spaces.

Some tools used in dynamic systems theory to describe complex systems are also useful to study cognition, such as trajectories running through state space. With these tools, one can model behaviour that is thought to be indicative of certain cognitive capacities (in section 6.2 below, we will consider some examples). If dynamic systems theory is thus applied, we can speak of dynamical cognitive science. In that sense, the resulting view of cognition, which we may refer to as dynamicism, can be viewed as yet another model of the mind (Van Gelder 1995, 1998), alongside (and to a certain degree, in competition with) classic symbolic computationalism and connectionism.

However, by viewing the mind, or perhaps more accurately, the cognizer, as a complex dynamic system, dynamicism does represent a break from the traditional view of cognition as presented in section 5.4. For all their rivalry, computationalism and connectionism shared the idea that cognition is to be understood in terms of internal representations, although they differed on the realization and the content of these representations. Dynamicism on the other hand promises a way to understand cognition without the need to invoke the concept of mental representations, namely by viewing cognition as a trajectory through state space (Van Gelder 1995; Port & Van Gelder 1995).

There are two further features of dynamical cognitive science that are worth mentioning at this point. First (as I already hinted at in section 5.1), whereas classic computationalism and connectionism can be said to study cognition within the boundaries of a preconceived framework, that is, as if it concerns processes going on inside an isolated subject, dynamicists stress the input of environmental factors in cognitive processes[56]. Feedback loops run from the organism to the environment, so that both co-evolve ('online cognition'). As I have already stated, the emerging picture of cognition seems to lend support to, or is at least compatible with, a movement in the philosophy of mind that rejects the internalist perspective, and instead views mind and cognition as extended, i.e. as arising from the interaction between the subject and the

[56] Strictly speaking, connectionists do acknowledge some outside influence, in the sense that connectionist networks receive input from the environment. This environment however, is very restricted and artificial – a far cry from the mutual feedback between environment, body and brain envisaged by dynamicism.

external world (Clark 1997, Clark & Chalmers 1998).[57] Second, because complex systems stand in this dynamic, feedback relation to the environment, with continuous mutual adaptation through time, they can raise problems regarding causality. Here, traditional, efficient causality (billiard balls striking each other, rocks shattering windows etc.) runs into problems. Instead, the relation between complex systems and their environment is one of *continuous reciprocal causation*: both are continuously affecting, and simultaneously being affected by, the other (Clark 1998 p. 356). In such systems, causality is no longer a relation between ordered pairs of events or objects, but instead is spread out between the neural, bodily and environmental domain (Wheeler & Clark 1999).

These features are important in the present context because they have repercussions for the explanations furnished by dynamical cognitive scientists. To appreciate this, let us briefly consider an example, borrowed from Wheeler and Clark (1999 pp. 106-108). If a new-born human infant is held in an upright position it will produce coordinated stepping movements. This lasts about two months – after that, the phenomenon disappears, until it reappears when the infant is 8 to 10 months old. Why is this? Rather than referring to, for example, rule-guided learning processes, dynamic systems theorists Thelen and Smith explain this behaviour in terms of self-organisation of the system, where this system involves the infant's brain, body and environment (1994). These researchers showed that the environment was an important factor, by holding 7-month-old infants (the phase during which the locomotion is absent) in the upright position over a motorised, moving treadmill, whereupon the stepping motion reappeared. Thelen and Smith's hypothesis was that the treadmill produces the behaviour by providing a substitute for certain leg-dynamics that occur in normal locomotion. As soon as the infant's legs touch the treadmill, the trailing behind of one leg acts as a spring so that it swings forward automatically: "The initiation of the swing appears to be triggered by the proprioceptive, biochemical information available to the central nervous system at the point of maximum stretch..." Wheeler & Clark 1999p. 107).

[57] Although to my mind, the combination of dynamical cognitive science and extended cognition is an obvious one, there is a recent tendency to interpret dynamical cognitive science as describing extended *mechanisms* (Zednik 2011; Kaplan 2012). Of course, extended cognition and extended mechanisms are not incompatible; rather, the point seems to be that the simple claim that cognitive processes span brain, body and environment, fails to pick out, from the multitude of factors that are causally relevant for cognitive processes, those factors that are actually part of the mechanisms responsible for those processes, so as to distinguish them from the causal factors that are more aptly described as causal background conditions (Kaplan 2012 p. 568). In this respect, it seems that focusing on how one can draw appropriate boundaries around an extended mechanism, as opposed to an extended cognitive process, could allow one to better understand the dynamic and reciprocal relations between the individual and the environment, as they are described by the models of dynamical cognitive science.

Thus, the environmental factors (the treadmill) act as a cue for the mature walking motion to self-organise in the body of the child, even though that particular skill (mature walking) is presumed lost at that particular stage of the child's development. Wheeler and Clark are quick to draw the lesson:

> ...in cases of on-line intelligence, no aspect of the extended brain-body-environment causal system should be explanatorily privileged in the cognitive-scientific understanding of the observed behaviour. This sharing-out of the explanatory weight – call it explanatory spread – is manifestly at odds with the cognitive-scientific tradition, which is to seek explanations of intelligent behaviour that appeal fundamentally, and usually exclusively, to strictly agent-internal neural phenomena (Wheeler & Clark 1999 p. 108).

The resulting picture in some respects resembles a mechanistic model. Recall that a mechanism has entities, activities and an organisation that combine to produce the overall behaviour of the system. Dynamic systems theory seems to provide a natural way of describing how behaviour results from the interaction of various entities carrying out various sub-routines: treadmills stretching muscles, biochemical stimulation of certain neural areas triggering the leg to swing forward, etc. Still, the models of dynamical systems theory are highly abstract and idealized, connecting mathematical variables rather than actual biological entities and activities. Moreover, as we shall see below, the model explanations furnished by dynamic systems theory have unmistakable CL-properties. In any case, when reflecting on the explanation of auto-locomotion in infants as Wheeler and Clark describe it, one important point emerges: the system described may be dynamic, complex and extend beyond the organism (the infant), *it is nevertheless a causal explanation*. The fact that this causality is spread out in this context simply means that the causes referred to in the explanation are not as unambiguously temporally and spatially located as traditional, representational explanations would have them. This causal feature will become important in the next section.

6.2 Two examples: rhythmic finger tapping and infant perseverative reaching[58]

In the previous section, I have mentioned several issues that are important to explanations in dynamical cognitive science: mathematical simulations, mechanistic models, prediction and causality. To see how these issues come together, I will consider two concrete examples taken from the literature on dynamical cognitive science itself.

Our first example is the HKB model of rhythmic finger tapping, named after its originators (Haken, Kelso & Bunz 1985), as interpreted in Kelso (1995). This model attempts to explain the curious observation that test subjects, having placed their hands palm-down on a table, can oscillate both index fingers in 'phase motion' (to the left and right at the same time) reliably across higher frequencies than they can oscillate them in 'antiphase motion' (to the left with one finger while to the right with the other). When Haken et al. asked their test subjects to increase the frequency of the antiphase movement, then upon reaching a critical speed, they would automatically switch to the in-phase mode. On the other hand, test subjects who performed the phase motion would not switch to antiphase motion. The explanatory question boils down to:

> Q1 Why can test subjects oscillate both index fingers in phase motion at much higher frequencies than in antiphase motion?

In terms of the HKB model, we have one collective variable, namely 'relative phase', and a differential equation describing how this variable changes as a result of a control parameter (frequency of oscillation) over time. Thus, Haken et al. are able not only to describe behaviour that has already been observed, but also to predict behaviour that can be (and indeed, has been) confirmed by subsequent experiments. Of philosophical importance is the fact that they do this without postulating an inner switching mechanism (homunculus or central executive deciding when to switch): the switch from two stable states to one is described as the result of the self-organising properties.

The second example is Thelen et al.'s model of infant perseverative reaching (Thelen & Smith 1994; Thelen et al. 2001). This model attempts to explain the 'A-not-B error', a phenomenon first described by Piaget (1954, 1963): a child between seven and twelve months old is presented with two boxes. When an adult comes in and hides a toy or piece of candy under one of the boxes, the child will reach for the correct box. Yet if the

[58] Parts of this and the succeeding section contains material published in Gervais & Weber (2011).

adult repeats this procedure several times and then suddenly hides the toy under the other box, the child will still reach for the first box, even though it has observed the adult hiding the toy under the other box. This gives rise to the following explanation-seeking question:

Q2 Why do infants between 7 to 12 months old make the A-not-B error, while older children do not?

Since its discovery, many explanations have been proposed to answer this question. Piaget's own explanation was in terms of object permanence, the idea being the children between 7 to 12 months old make the error because they lack the understanding that objects continue to exist despite not being observable any longer. Other researchers explain the phenomenon in terms of information processing, interpreting the difference between the age groups as having to do with the amount of cognitive resources available, in particular memory, attention and planning (Diamond 1985). Connectionists describe the error as resulting from the competition between latent memory traces for A and active traces for B. In infants between 7 to 12 months old, the latent traces 'win' the competition, while older infants are better at actively maintaining memory traces, so that they do not make the error (Munkata 1998).

What these explanations have in common is that they all explain the occurrence of the error in terms of the infant's internal cognitive processes. They refer to generalized statements about mental representations the infants (supposedly) have and the computational processes acting on these representations. In effect, they are of the inferential type described in section 5.4. In contrast, the dynamical systems model focuses on the reaching behaviour of the infants, quantifying all the relevant elements and describing these by means of a mathematical equation.

6.3 Walmsley's analysis evaluated: non-deductive causal CL explanations

There are important similarities between the two dynamical explanations considered in the previous section. Walmsley notes (2008) that a particular pattern of behaviour in this kind of experimental setup follows as a mathematical and deductive consequence of the equation in conjunction with the initial states, and has the same logical form as the prediction of that event would have taken. Reflecting on the features of these explanations (explanandum as a logical consequence of the explanans, equivalence of

explanation and prediction), Walmsley draws the conclusion that "...some dynamical models provide covering law explanations" (2008, p. 342). In the context of the present discussion, this claim is of course very interesting, so let us take a moment to evaluate it.

It is undeniably true that the characteristics Walmsley mentions are reminiscent of Hempel's covering law model. To put it briefly, it seems that these dynamical systems models involve explanation by subsumption. However, from this it does not follow that they are covering law explanations in a strict Hempelian sense. More specifically, there are two problems with Walmsley's interpretation:

(1) By characterizing explanations in dynamical cognitive science as deductive-nomological Walmsley neglects an important property of the explanations, viz. that they are causal. In this way he suggests that they are problematic, while they are not.

(2) Not all explanations in dynamical cognitive science are deductive-nomological. At least some of them fit a non-deductive variant of the covering law model, which uses so-called default rules instead of strict, exceptionless laws.

That is to say, although I believe Walmsley's analysis puts us on the right track, we need to add some modifications to understand what is going on in these dynamical systems models.

As we may recall from our discussion in sections 2.2.3 and 2.2.4, Hempel's model has problems with the asymmetry of explanation[59], and Hausman offers an obvious solution by introducing a causal requirement. To make sense of the dynamical systems explanations we should take Hausman's lesson seriously. Of course, Hausman is able to avoid the flagpole counterexample. More importantly for our present purposes however, are the *pragmatic* benefits that accepting Hausman's causal requirement brings with it. In the special sciences particularly, we typically want not only to explain phenomena, but also to control and manipulate them. In dynamical cognitive science, we are interested in explanations because they enable us to control and manipulate behaviour to e.g. enhance our educational methods, to develop new drugs etc.[60] If we desire to put dynamical cognitive science to good use for humanity, then we should

[59] To be fair, Walmsley does mention a number of the classic counterexamples against Hempel's model in notes 33 and 34 of his 2008 article

[60] This point extends beyond dynamical cognitive science to all special sciences. Here is a quote from Carl Craver making the same point about neuroscience: "Neuroscience is driven by two goals. One goal [...] is explanation ... The second goal of neuroscience is to control the brain and the central nervous system. Neuroscience is driven in large part by the desire to diagnose and treat diseases, to repair brain damage, to enhance brain function, and to prevent the brain's decay" (2007, p. 1).

make sure that these explanations pick out the real causes; i.e. allow us to manipulate, control and predict.

The explanations presented by the authors discussed by Walmsley, satisfy this additional criterion. Thelen et al.'s model of infant perseverative reaching explains an infant's reaching behaviour as a mathematical consequence of an equation with a number of values for the parameters and variables, including the current state of the movement field, the general and specific memory inputs to the system and a function integrating competing inputs. From the perspective of the traditional CL model, there is nothing against reversing the order of the argument: from the equation, parameters and variables, together with the reaching behaviour of the child, we might deduce (hence 'explain') under which box the researcher has hidden the toy. Surely, this kind of argument has to be ruled out as an explanation. And what is important: Thelen et al. do not claim that such arguments are explanations. They only claim that the argument which starts from the causes and has the effect as conclusion is an explanation. Similarly, the authors of the HKB model do not claim that effects can explain their causes. Of course, no one expects practicing scientists to explicitly adhere to one or other model of explanation. Like Walmsley himself, I am simply trying to reconstruct what format their explanations assume, and it is in this context that their silence on this issue is relevant, although it does not count as evidence.

The upshot of this is that we can give a more precise characterisation of the examples: they fit the "causal-deductive-nomological model" of Hausman. This is important because, if they would not fit this model, dynamical cognitive science could have been accused of providing pseudo-explanations like the one in which the height of a flagpole is explained by the length of its shadow, instead of genuine explanations. Walmsley acknowledges the validity of the counterexamples raised against the traditional covering law model. In his view they "...show [...] that explanations in dynamical cognitive science will be the subject to the same set of criticisms, in virtue of the form they take" (2008, p. 344). This conclusion is not correct: the explanations have a more specific form than Walmsley admits, and therefore do *not* share the problems of Hempel's model.

As we have seen in section 2.2.2, in his 1965 Hempel distinguished between DN and IS explanations. Walmsley does not consider the latter type of explanations. He only discusses Hempel's DN model (2008, p. 338-340) and claims that the explanations of dynamical cognitive science conform to this model. He explicitly says that in the explanations of Thelen et al. and in the HKB explanations, the explanandum is a *deductive* consequence of the explanans (pp. 340-341). This is strange, because his paper also contains information that seems to contradict this conclusion. When describing the effect explained by Thelen et al., he mentions that it is "...enormously sensitive to slight changes in the experimental conditions, such as the delay between viewing and reaching, the way the scene is viewed, the number of trials, the presence of distracting

stimuli, and so on" (p. 335). The model of Thelen et al. is superior to previous models because it can take into account many of these contextual subtleties. However, as long as the model cannot cope with *all* contextual variation, it cannot produce deductive-nomological explanations. The reason is that, if the model cannot account for all contextual variation, the law we can derive from it has the form 'If initial conditions C_1, C_2, ... C_N are satisfied, then *usually* E happens'. In order to have DN explanations however, we need 'always' in the law instead of 'usually': DN explanations need strict, exceptionless laws of the form 'If initial conditions C_1, C_2, ... C_N are satisfied, then E *always* happens'.

The law we actually can derive from Thelen et al.'s model has the format of a default rule. Default rules (e.g. 'Birds usually fly') differ from universal generalizations in that they allow exceptions (e.g. 'Penguins don't fly'). Default rules also differ from probability statements in that they do not specify the relative frequency of the exceptions and 'normal' cases ('usually' can mean anything fairly close to probability 1). Another important characteristic of default rules is that we can formulate them without knowing where the exceptions lie (we can say 'birds usually fly' and have good evidence for that claim without knowing which species or individuals are the exceptions).

In my view the best way to characterize the explanations given by Thelen et al. is: they are non-deductive causal CL explanations using a default rule. They are causal, as established in the previous section. They use a covering law and have the form of an argument (this has been shown by Walmsley). Yet the covering laws they use admit of exceptions, so the explanations are not deductive. This means that they either use a default rule (as suggested here) or a probability statement (and thus would be what Hempel calls inductive-statistical explanations). The latter presuppose that we can give a precise relative frequency of 'normal cases' and 'exceptions'. This does not seem the case in the explanations of Thelen et al., so they should be seen as explanations using default rules. As Walmsley himself acknowledges (2008, p. 344), exceptionless generalisations have been hard to find in psychology. So there are reasons to suppose that explanations using default rules are the rule, rather than the exception.

Chapter 7
CL explanations and mechanistic models in biology

7.1 Introducing biology: some assumptions and the chapter outline[61]

Although a general introduction to biology would take far too much room, if only for the fact that like cognitive science, biology is a highly heterogeneous discipline, comprised of sub-disciplines such as ecology, population dynamics, genetics, cell biology and molecular biology, it is nevertheless important to highlight two assumptions that are important when considering mechanistic explanations in the context of biology. Although I offer these assumptions in this particular context, they apply no less to cognitive science, and indeed we have already encountered them in previous chapters.

First, just as we have seen in the case of cognitive science, capacities constitute a major type of explanandum in biology. Indeed, the biological capacities that stand in need of explanation are many. How do cells reproduce? How do plants and bacteria convert carbon dioxide into organic compounds? How are genetic traits preserved through generations? How do humans see depth? Second, like their cognitive counterparts, biological capacities are often explained by means of a mechanistic explanation. Recall that a mechanistic explanation of a capacity is an explanation that explains the capacity by means of a description or model of a mechanism, where mechanisms, to use Machamer, Darden and Craver's often quoted definition again, are "...entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (2000, p. 3). Because of these

[61] Part of this chapter is taken from Gervais & Weber (2013b).

characteristics, many philosophers have rightly stressed the importance of mechanisms and mechanistic explanations in biology (e.g. Woodward 2001, 2002; Glennan 2002, 2005; Baker 2005; Darden 2005, 2006; Bechtel 2006, Bechtel & Abrahamsen 2005).

Second, ontologically speaking, we may safely assume that every biological capacity has *a* mechanism responsible for it: there are no 'free floating' functions or capacities. Epistemologically speaking however, our knowledge of these mechanisms is likely to be incomplete to a lesser or greater degree. This situation has led to the idea that explanatory progress in biology consists of filling in more and more details of our models as they become known to us. On this account, it is possible to rank mechanistic explanations on a continuum, ranging from how-possibly models, where the model is still abstract and highly speculative, via how-plausibly models, to how-actually models, which constitute a complete and accurate description of the mechanism responsible for a given biological capacity (Craver 2006).[62] Of course, how-actually models are more a regulative ideal than that they are realistically attainable, but the general idea is that as one adds flesh to a skeletal model, its explanatory value increases (cf. Piccinini & Craver 2011).

However, what about those situations in which biologists know *next to nothing* about the mechanism responsible for a given capacity? In such a case, in keeping with the ontological assumption mentioned above, all we know is that there must be *a* mechanism, although its entities, activities and organization are unknown. As I will show, this is by no means an uncommon situation. The continuum-position described above, plausible as it may seem, suggests that in such cases, all we can do is wait until enough of these details become known – only then will we have any explanatory hold over the capacity in question.

Yet, as I will argue below, this does not accurately reflect what is going on in biology. In fact, capacities whose underlying mechanisms are unknown can and do figure in genuine biological explanations. Since these explanations neither make reference to the entities nor the activities of the mechanism, they are non-mechanistic. Hence, it seems that here, the conceptual apparatus of mechanistic explanations is not suitable to understand the explanatory practices of biologists. Rather, to understand what is going on in these situations, we must turn to the DN model instead.

As we shall see in the next chapter, this is somewhat going against the current of present day philosophy of biology, and I immediately concede that the DN model should not be understood in the strict, Hempelian sense of rigidly deducing the explanandum

[62] I will consider Craver's continuum in more detail in chapter 10.

from a universal law of nature in combination with boundary conditions.[63] Also, I want to make it clear that I am not *against* mechanistic explanations. Rather, I defend the complimentary thesis that besides mechanistic explanation, *a certain type of non-mechanistic DN explanation is crucial to biology*. Only when this is granted, can we make sense of the type of explanatory situation sketched above. That is, I believe that the current debate about biological explanation suffers from a misleading dichotomy between DN explanations on the one hand, and mechanistic explanations on the other. In a sense, what I advocate is a third alternative: some explanations in biology deduce the explanandum from generalizations that contain only very limited information about the underlying mechanisms. Although the generalizations involved are highly contingent, and the semantic requirements for these explanations take them well beyond Hempel's original account, formally speaking, they adhere to the DN model. At the close of this chapter, I will argue that the mechanists' dismissal of the DN model in fact throws out the baby with the bathwater. To be sure, the mechanists are right in their assessment that the *semantic* requirements of Hempel's model are not suitable in the field of biological explanation, but it is wrong to conclude from this that the *formal* requirements (deduction of the explanandum from premises containing at least one generality, together with boundary conditions) of the DN model are never met, or are only met in contrived, artificial situations. As I will show over the course of this chapter, there are many scientifically interesting cases which adhere to the DN standards as we shall construe them, and thus are illustrative of the third alternative I shall propose.

Let me conclude this introduction with an overview of the chapter. In Sections 7.2 and 7.3 I will present two examples of situations in which biological capacities (pigeon navigation and photoperiodism respectively) are explained while the underlying mechanism is largely unknown. As I will show, these explanations take a basic DN format. Reflecting on these examples, in Section 7.4 I argue that explanations of this type are crucial, for three reasons: they are sometimes the only thing we have, they are heuristically useful, and they provide understanding. In Section 7.5 I discuss the relation between the types of explanations we distinguish and the types of experiments biologists often perform in order to gather evidence for and against explanations. Here, I will argue that the explanations discussed in sections 7.2 and 7.3 are backed up by hypothetico-deductive experiments.

---

[63] One of the main concerns with the DN model in the context of biology is the status of biological laws. Laws, in the sense of strict, exceptionless, non-contingent regularities, seem (largely) absent from biology (Beatty 1995, 1997; Brandon 1997; Sober 1997). In section 8.2 I will consider this issue in somewhat more detail; for the moment however I am content to think of biological laws in a revised sense, along the lines of Mitchell (1997, 2000), according to which they are pragmatic generalizations that allow for prediction, explanation and manipulation, whether they succeed or fail to meet the traditional criteria for lawhood.

7.2 Pigeon navigation

7.2.1 Possible DN explanations of pigeon navigation

The first set of examples of DN explanations in biology I shall consider relates to the capacity of homing pigeons (*Columba livia*) to navigate. The source for this example is Keeton & Gould 1986, pp. 575-585. A trained pigeon can be taken from home, transported over very long distances (hundreds of miles are not uncommon) and still find the way back to its home after being released – a phenomenon that has been known since antiquity. Moreover, it appears that pigeons are able to exercise this capacity both in sunny weather and on cloudy days. This gives rise to two explananda:

E1 Pigeons have the capacity to find their way back home on sunny days.
E2 Pigeons have the capacity to find their way back home on clouded days.

With respect to the first explanandum, it was shown that that pigeon navigation depends on the position of the sun as a reference point. Thus, a DN explanation of E1 posits an internal sun compass:

L1 Pigeons have a sun compass.
L2 All animals with a sun compass have the capacity to find their way back home on sunny days.

E1 Pigeons have the capacity to find their way back home on sunny days

As I have already mentioned above, pigeons that are released on a clouded day also have the capacity to find the way back home, hence E2. In a set of experiments, W. T. Keeton demonstrated that pigeons also have a magnetic compass. Keeton released birds that had magnets and birds with brass rods (which function as placebos) attached to them, both on sunny and on clouded days. The results were clear: on sunny days, the birds were unaffected, but on clouded days the birds carrying magnets became disoriented. In this way, Keeton arrived at the following DN explanation for E2:

L3 Pigeons have a magnetic compass.

L4 All animals with a magnetic compass have the capacity to find the way back home on clouded days.

E2 Pigeons have the capacity to find the way back home on clouded days.

Furthermore, the experiments of Keeton suggest conditions under which the two systems become operative:

L5 Pigeons have a sun compass and a backup magnetic compass that works only on cloudy days.

L6 All animals with a sun compass and a magnetic compass that works only on cloudy days, have the capacity to find the way back to their house on sunny days even if they carry a magnet around their neck.

E3 Pigeons have the capacity to find the way back to their house on sunny days, even if the carry a magnet around their neck.

In short, the idea is that if the sun compass cannot be used, the magnetic compass comes into play as a kind of backup system.

7.2.2 But what are the mechanisms?

In the previous section I presented three examples of explanations of biological capacities. Let us now analyse their properties. First, they are DN explanations, which means that DN explanations are at least possible in biology. Second, they posit the existence of a mechanism without describing it. Let me clarify what I mean by this. The claim that pigeons have a solar compass is, in my view, identical in meaning to the following claim:

> In the body of pigeons there are entities (of which we don't know where they are and what they look like) that have certain unknown activities and are organized in an unknown way. These entities, activities and organization ensure that pigeons have the capacity (on sunny days) to determine the angle they have to maintain relative to sun.

The claim that pigeons have a magnetic compass is in my view identical in meaning to the following:

> In the body of pigeons there are entities (of which we don't know where they are and what the look like) that have certain unknown activities and are organized in an unknown way. These entities, activities and organization ensure that pigeons have the capacity (on clouded days) to determine the angle they have to maintain relative to the magnetic field of the earth.

If one agrees that this is the meaning of the laws L1 and L3, the explanations are non-mechanistic (because no information is given about the entities, activities or organization). However, from an ontological point of view they presuppose a mechanism: the law cannot be true unless there is *a* mechanism.

Of course, the mechanists can maintain that these are not examples of genuine explanations – that is, they can maintain the 'mere description' objection that is frequently raised against the CL model. In Section 7.4 therefore, I will provide some arguments to show that explanations of this type are valuable for a number of reasons. First however, let us look at a second set of examples.

7.3 Photoperiodism

7.3.1 Possible CL explanations of photoperiodism

For the second set of examples, let us turn to W. W. Garner and H. A. Allard's explanation of the capacity of plants to flower. My sources are Keeton & Gould 1986, pp. 395-402 and Murneek 1948. Although the capacity to flower is a crucial part of the reproduction cycle of heterosporous plants, the details of the process vary widely across different species. For example, some flower in the spring, others in the summer or autumn.

Garner and Allard noticed that a new variety of tobacco plant, the Maryland mammoth variety (*Nicotiana tabacum*), grew to excessive heights in the summer without blooming. However, if they took cuttings of the plant and grew them in the greenhouse during winter, the plant *would* bloom. Thus, the following explanandum presented itself:

E4 Tobacco plants of the Maryland mammoth variety have the capacity to flower in the greenhouse in the winter.

In addressing this explanandum, Garner and Allard conducted a series of experiments, eliminating as many variables as possible, and found that the apparent key determining

factor was the number of hours of daylight the plants received. With this information, they could induce flowering in plants during the summer if they shielded the plants from sunlight for part of the day, while in the greenhouse they could inhibit blooming in the winter by exposing them to artificial lights during the night. Based on further research, they distinguished between three groups of plants: short-day plants, which flower if there is a relatively short daily exposure to light (usually less than 12 to 14 hours); long-day plants, which flower if there is a relatively long daily exposure to light (usually more than 12 to 14 hours), and day-neutral plants (plants in which flowering is not influenced by length of daily exposure to light). They dubbed this link between the length of exposure to light and flowering *photoperiodism*. The first results were published in the *Journal of Agricultural Research* in 1920, and they published several papers on this topic during the next 25 years.

Subsequent research revealed that it is actually not the length of the day, but the length of the night which determines whether plants flower or not. Studies showed that it is not possible to prevent a long-day plant from flowering at the proper season by shielding it from light for an hour during the middle of the day. However, it is possible to prevent a short-day plant to flower in season by exposing it to light for a short interval during the night (minutes or even seconds). Similarly, long-day plants can be caused to flower in the wrong season (i.e. during winter) by means of a light flash that interrupts the dark period. In short, the key factor in the flowering of Maryland mammoth plants is the amount of time the surface of the leaves are *not exposed to sunlight*. However, the terminology (long-day, short-day, day-neutral) has been preserved, with a different meaning. A short day plant is a species that flowers after dark periods of a minimum duration. A long day plant is a species that requires short dark periods in order to flower. This gives us the ingredients to construct a possible DN explanation for E4:

L7 Tobacco plants of the Maryland mammoth are short-day plants with a minimal dark-period-length of 12[64] hours

A1 In the greenhouse during winter the amount of time the plants are not exposed to light exceeds 12 hours.

E4 Tobacco plants of the Maryland mammoth variety have the capacity to flower in the greenhouse in the winter.

In this explanation, L7 is a law and A1 is an auxiliary hypothesis.

It is easy to see that similar explanations can be given for the behaviour of other short-day plants. This is exactly what Garner and Allard did in the case of the Biloxi soybeans (*Glycine max*), which they planted throughout the months of May, June and July. Even though these intervals meant that there were considerable differences in the growing periods of the soybeans, they nevertheless began to flower in September en masse, giving rise to the explanation:

L8 Biloxi soybeans are short-day plants with a minimal dark-period-length of 10 hours.

A2 From September on the amount of time the plants are not exposed to light exceeds 10 hours.

E5 Biloxi Soybeans flower in September, regardless of the month in which they are planted.

With the division between the three plant groups in place, one can make similar explanations of the behaviour of other plants, such as chrysanthemum, dahlia and cocklebur (short-day) and beet, clover and larkspur (long-day).

[64] It should be kept in mind that the numbers cited in this section are *approximate* values. In their 1920 article, Garner and Allard themselves drew the somewhat rough distinction between short- and long-day plants that I mentioned above, but in their summary they did note that: "In a number of species studied it has been found that normally the plant can attain the flowering and fruiting stages only when the length of day falls within certain limits..." (Garner & Allard 1920 p. 603). Subsequent research must of course be carried out to give more precise values for each species. Despite the fact that the Garner and Allard's experiments are well documented, the particular values for the Maryland mammoth and the Biloxi soy bean are not easily obtainable. For the Maryland mammoth, I rely on Foster and Kreitzman 2009; in the case of the Biloxi soy bean, on Garner 1933 (p. 349). This gives us minimal dark-period-lengths of 12 and 10 hours respectively. Although these values are approximate, and subsequent research may yield different/more precise ones, this will make no difference to the philosophical points developed in this section.

7.3.2 But again, what are the mechanisms?

Let us now analyse the examples of section 7.3.1. There is an important difference with the examples of the homing capacity of pigeons: the two explanations of the previous section *do not posit any mechanism.* They are DN explanations, but not of the variety I consider interesting. Of course, this is chiefly due to the way I have framed them; to be sure, Garner and Allard themselves would never have *denied* that there is a mechanism responsible for photoperiodism. However, I have given these examples for two reasons. The first is heuristic: they introduce the scientific material necessary to understand my next set of examples, which *do* fit the criterion (they posit a mechanism without describing it). Second, they show that it is *possible* to construct DN explanations of which the explanatory value is debatable, exactly because all they do is subsume the capacity under a more general regularity. Indeed, one may argue (and a mechanicist would certainly concur with this claim) that the examples of the previous section are good predictive arguments but no explanations. That view is compatible with what I defend in this chapter (I will not explore the explanatory value of such arguments further).

Let us now take a further step. In 1936 the Russian researcher M. H. Chailakhian removed the leaves from the upper half of chrysanthemums and put paper between the upper and lower half, so that one could be exposed to daylight while the other remained shielded. Next, he exposed this upper half to long days, and the untreated lower half to short days, which resulted in the plant flowering. The reversed procedure, exposing the lower half to long days and the defoliated upper half to short days, resulted in the plant not flowering. From these experiments, Chailakhian concluded that the length of day influenced flower buds only indirectly, through causing the leaves to produce a hormone that induced the buds to flower. This hormone however, which he called *florigen*, was hypothetical, and as such the mechanism he proposed was at best speculative. Indeed, more than seventy years of subsequent research failed to isolate this hypothesized hormone and describe its chemical structure. This has led some biologists to the conclusion that there is no florigen, or that the term at best refers to a functionally defined concept, the actual filler of which still needs to be identified. To give an example of this latter type of undertaking, one proposal was that flowering would be controlled by the ratio of two or more other hormones.

Whatever the exact mechanism is, the experiments of Chailakhian show that light does not stimulate flowering by acting directly on the buds: there is a stimulus (whatever it is) that is passed on from the leaves of the plant to the buds. This insight allows us to explain the capacities E4 and E5 by means of an explanation of the type we want. For E4 we have:

L9 Tobacco plants of the Maryland mammoth are short-day plants which start to produce and transport the flower stimulus if its leaves are exposed to a minimal dark period of 12 hours.

A3 In the greenhouse during winter the amount of time the leaves of the plants are not exposed to light exceeds 12 hours.

E4 Tobacco plants of the Maryland mammoth variety have the capacity to flower in the greenhouse in the winter.

Note that there is a small but crucial difference between A3 and A1: the new auxiliary hypothesis is about the leaves of the plant, rather than about the plant as a whole (for example, the roots are never mentioned in the explanation).

For E5 we have:

L10 Biloxi soybeans are short-day plants which start to produce and transport the flower stimulus if its leaves are exposed to a minimal dark period of 10 hours.

A4 From September on the amount of time the leaves of the plants are not exposed to light exceeds 10 hours.

E5 Biloxi Soybeans flower in September, regardless of the month in which they are planted.

Like the examples of explanations targeting pigeon navigation, these two examples are DN explanations, and hence constitute a second set of examples which show that DN explanations are possible in biology. The explanations are of the same type as those in 7.2.1, in that they posit a mechanism (a flower stimulus responsible for the capacity to flower) without describing it. The explanations are non-mechanistic (because no information is given about the entities, activities or organization). However, from an ontological point of view they presuppose a mechanism: the law cannot be true unless there is one or more underlying mechanism.

7.4 The value of DN explanations in biology

Let us take stock. The two case studies support the following claims:

> (1) Biologists do construct DN explanations of capacities.
> (2) Some of these DN explanations have a specific structure: they posit a mechanism without describing it.

Presumably, claim (1) is not very controversial. Indeed, the mechanists can agree with both statements yet maintain that these explanations are uninteresting. In this section I will present three arguments against such a view: I argue that (i) these explanations are sometimes the only ones we have, even after many decades of research (7.4.1), (ii) that they are heuristically useful (7.4.2) and (iii) that they provide understanding[65] and thus are interesting in their own right (7.4.3).

7.4.1 The only thing we have

With respect to photoperiodism, I already mentioned that the mechanism is still subject to debate. Currently, the flowering is thought to result from the combination of photoreceptor proteins like phytochrome and cryptochrome and the circadian clock. Yet the circadian clock as such is no more than a place-holder term referring to some biochemical mechanism, which receives environmental cues as input and has certain behaviour as output. In this highly abstract form, it is postulated to explain the circadian rhythms of plants and fungi as well as animals. The precise details of this clock will vary from species to species, even within the realm of plants. To arrive at a full mechanistic explanation for photoperiodism in the Maryland mammoth and Biloxi soybean, more work is needed.

The same applies to the homing pigeons. Though the experiments of Keeton date from the 1970s (they were published in *Scientific American* in 1974*)* it is still impossible to give mechanistic explanations of the capacities. To illustrate this, let us take a closer look at the magnetic compass. A relatively recent proposal is that iron particles (superparamagnetic magnetite or SPM particles) in the nerve terminals of sensory nerves in the upper beak of the homing pigeons might play a role in the mechanism

[65] Usually, the term 'understanding' is used by philosophers of science to denote whatever it is that scientists seek to increase or achieve by constructing explanations. In what follows, we shall employ the term in this loose, non-technical sense, and allow the criteria for understanding as used by scientists, to be subject to contextual (e.g. historical) variation (De Regt & Dieks 2005).

underlying their capacity to find the way home (Fleissner et al. 2003). The hypothesis is that these particles react to the magnetic field of the earth, so that the nerve cells in the upper beak act as magnetoreceptors, passing on information to the brain, allowing the pigeon to determine its direction, height and location. However, the researchers stress that these SPM particle-clusters in the beak are only a *candidate* for the responsible magnetoreceptor (Fleissner et al. 2003 p. 360). Moreover, their conclusions have been disputed recently by a group of researchers who argue that the cells in which the SPM particles are found, are in fact not nerve cells at all, but rather specialized white blood cells (macrophages), whose function is to recycle iron particles of red blood cells (Treiber et al. 2012).[66] If that is true, then it is implausible that they play a role as magnetoreceptors, as white blood cells do not possess the ability to convey information to the brain. Furthermore, the number of SPM cells varies widely among individual pigeon beaks, which seems at odds with their supposed roles as magnetoreceptors. Treiber et al. conclude that "...our work reveals that the sensory cells that are responsible for trigeminally mediated magnetic sensation in birds remain undiscovered. These enigmatic cells may reside in the olfactory epithelium, a sensory structure that has been implicated in magnetoreception in the rainbow trout" (Treiber et al. 2012 p. 369). In short, the message of Treiber et al. is: 'we don't know yet where the magnetic compass is located, and since we have failed with the beak, let's now look at the nasal cavity, because that is where rainbow trout have it'. Interestingly, the next hypothesis that will be considered by this research group apparently results from a comparatively simple instance of analogy reasoning. Borrowing some terminology from Machamer et al., this does not even amount to a 'mechanism sketch' (2000 p. 18), let alone a full blown mechanistic explanation.

7.4.2 The heuristic value of CL explanations in biology

The heuristic value of CL explanations that posit but do not describe mechanisms lies in the fact that they suggest new explananda. In the pigeon case, the explanations lead to the following new explananda

E6 Why do pigeons have the capacity (on sunny days) to determine the angle they have to maintain relative to the sun?

E7 Why do pigeons have the capacity (on clouded days) to determine the angle they have to maintain relative to the magnetic field of the earth?

[66] See the cover image of this dissertation.

If these questions are answered, the results can be used to build a mechanistic explanation for the original explananda (E1 and E2). In other words: the explanations we have considered here are useful steps towards mechanistic explanation. In effect, they share this virtue (of suggesting new explananda) with mechanistic explanations. This is also the case in the photoperiodism example, were the explanations in 7.3 lead to the following questions:

E8 Why do plants have the capacity to produce a flowering stimulus in their leaves which depends on night length?

E9 Why do plants have the capacity to transmit this flowering stimulus from the leaves to the bud?

7.4.3 The intrinsic value of DN explanations in biology

Finally, I think that the type of DN explanations we have been considering (i.e. the ones that posit mechanisms without describing them) are intrinsically valuable because they provide understanding. More precisely, they allow us to understand *contrasts.* Let us go back to the pigeons one more time. Consider the following question:

E10 Why do pigeons have the capacity to find their way back home while other sedentary birds do not have this capacity?

The explanations discussed in Section 3 suffice to understand this contrast: the other sedentary birds don't have a solar compass, nor a magnetic one. We do not need the details about how the capacity is implemented in pigeons in order to understand what makes pigeons special compared to other species of resident birds.

Another example is this

E11 Why do woodcocks migrate during the night, while pigeons cover long distances during the day?

Here the answer is that woodcocks, like other nocturnal migrants, use stellar constellations as navigation cue. Again, we do not need to know the details in order to understand the difference between the two species.

7.5 Types of explanations and types of experiments

We have seen that a certain type of DN explanation exists in biology, and that they are valuable for a number of reasons. However, there is a further argument in favour of DN explanations that posit but do not describe a mechanism: they are supported by hypothetico-deductive experiments. In this section, I discuss the relation between different types of explanation, and the types of experiments biologists often perform in order to gather evidence for and against explanations.

The main challenge for scientists who want to give a mechanistic explanation of a capacity is to establish the constitutive relevance of the entities and activities that are mentioned in the explanans. Craver formulates this problem as follows:

> Not all parts are components. Consider again the difference between mechanisms and machines. Machines contain many parts that are not in any mechanism. The hubcaps, mud-flaps, and the windshield are all parts of the automobile, but they are not part of the mechanism that makes it run. They are not *relevant* parts of that mechanism. Good mechanistic explanatory texts describe all of the relevant components and their interactions, and they include none of the irrelevant components and interactions. (2007, p.140)

The crucial question is: how can we show that an entity X and its activity ϕ indeed are components of the mechanism of the ψ-ing of an S? Craver's answer is labelled the *mutual manipulability account*[67]:

> [A] component is relevant to the behavior of a mechanism as a whole when one can wiggle the behavior of the whole by wiggling the behavior of the component and one can wiggle the behavior of the component by wiggling the behavior as a whole. The two are related as part to whole and they are *mutually manipulable*. More formally:(i) X is part of S; (ii) in the conditions relevant to the request for explanation there is some change to X's ϕ-ing that changes S's ψ-ing; and (iii) in the conditions relevant to the request for explanation there is some change to S's ψ-ing that changes X's ϕ-ing. (2007, p. 153, italics in original)

Whether condition (ii) is satisfied can be tested by means of bottom-up experiments. Craver distinguishes two types: interference experiments (inhibitory bottom-up experiments) and stimulation experiments (excitatory bottom-up experiments). Here is Craver's characterization and example of the first subtype:

[67] For an up to date discussion of this account, see Leuridan 2012.

> In interference experiments, one intervenes to diminish, disable, or destroy some putative component in a lower-level mechanism and then detects the results of this intervention for the *explanandum phenomenon.* The assumption is that if X's ϕ-ing is a component in S's ψ-ing, then removing X or preventing it from ϕ-ing should have some effect on S's ability to ψ.
>
> Lesion experiments, for example, are interference experiments in which something intervenes to remove a portion of the brain and one then detects the effects of the lesion on task performance. (2007, p. 147)

The second subtype is characterized and illustrated follows:

> In stimulation experiments, one intervenes to excite or intensify some component in a mechanism and then detects the effects of that intervention on the *explanandum phenomenon.* The assumption is that if X's ϕ-ing is a component in S's ψ-ing, then one should be able to change or produce S's ψ-ing by stimulating X.
>
> The classic example of stimulation experiments is Gustav Fritsch and Eduard Hitzig's (1870) work on the motor cortex (see Bechtel forthcoming). Fritsch and Hitzig performed a series of experiments on dogs in which they delivered low-grade electrical stimuli to a cortical area now known as the motor strip (see Bechtel forthcoming). Localized stimuli along this area produce regular and repeatable movements in specific muscles, including the legs, the tail, and the facial muscles. (2007, p. 149)

Whether condition (iii) is satisfied can be tested by activation experiments, which are excitatory top-down experiments. Here are Craver's characterization and example:

> In activation experiments, one intervenes to activate, trigger, or augment the *explanandum phenomenon* and then detects the properties or activities of one or more putative components of its mechanism. ... The basic assumption behind activation experiments is that if X is a component in S's ψ-ing, then there should be some difference in X depending on whether S is ψ-ing or not.
> ...
> There are several common varieties of activation experiment at all levels in neuroscience. In PET and fMRI studies, one activates a cognitive system by engaging the experimental subject in some task while monitoring the brain for markers of activity, such as blood flow or changes in oxygenation. (2007, p. 151)

The link between the experiments that Craver describes (and for which he uses the general label "interlevel experiments") and mechanistic explanations of capacities can be formulated in two ways:

(1) Biologists who want to give mechanistic explanations of capacities must do bottom-up and top-down experiments.

(2) The fact that biologists want to give mechanistic explanations of capacities explains why they do perform bottom-up and top-down experiments.

These are two sides of the same coin.

Bottom-up experiments and top-down experiments require that one intervenes on parts that may be components of the mechanism. This is impossible in cases where biologists give DN explanations of the type we are considering, because the parts cannot be localized. The information these experiments would yield is also superfluous: I don't make any claims about parts being components of the relevant mechanism or not; constitutive relevance is not an issue in these explanations.

How are DN explanations that posit yet do not describe mechanisms backed up? The experiments I have briefly described when discussing the explanations, are *hypothetico-deductive* experiments along the lines Hempel set out in his *Philosophy of Natural Science* (1966, ch. 2). From the hypothesis that a mechanism of a certain type is present, the scientists derive that a specific causal relation is to be expected. From the alternative hypothesis (the absence of the mechanism) they derive that the causal relation is expected to be absent. These hypothetical derivations have the following general format:

> If mechanism M is present, then one expects a causal relation between variable C and variable E.
>
> If mechanism M is absent, then one expects no causal relation between variable C and variable E.

Note that the derivations have causal relations in their consequent. The experiment is performed in order to find out whether or not the causal relation obtains. Let me illustrate this. In the experiments with the magnets, the hypothetical derivations are the following:

> If pigeons have a magnetic backup compass then on clouded days one expects a causal relation between carrying a magnet around the neck or not (C) and average flight direction (E).
>
> If pigeons do not have a magnetic backup compass then on clouded days one expects no causal relation between carrying a magnet around the neck or not (C) and average flight direction (E).

The experiments of Keeton are randomized trials which test exactly the causal relation that is at stake here: there is an experimental group with magnets around their neck and a control group with copper rods (which function as placebos). The difference between the average flight direction of the experimental and control group was statistically significant, so his experiments support claim that there is a causal relation between C and E. By means of *modus tollens*, the hypothesis that pigeons do not have magnetic backup compass is rejected.

As with Craver's interlevel experiments, the link between the hypothetico-deductive experiments and DN explanations that posit yet do not describe mechanisms can be formulated in two ways:

(1) Biologists who want to give a DN explanation of capacities that posits but does not describe a mechanism must do hypothetico-deductive experiments.

(2) The fact that biologists want to give such explanations of capacities explains why they do perform hypothetico-deductive experiments.

Again, these are two sides of the same coin.

Thus, we see how the type of DN-explanations we have been discussing in sections 7.2 and 7.3 are not only valuable (as I argued in 7.4), but are backed up by hypothetico-deductive experiments, just as the bottom-up and top-down experiments described by Craver support the use of mechanistic explanations. In both cases, if the type of explanation in question was not pursued by the scientists, then we would be hard pressed to explain why these scientists perform the types of experiments they do.

Chapter 8
Rethinking the dichotomy between CL explanations and mechanistic models in biology

8.1 State of the debate

The conclusion drawn at the end of section 6.3 was that dynamical cognitive science, as exemplified in HKB model of rhythmic finger tapping and Thelen et al.'s model of infant perseverative reaching, does contain a specific type of CL explanations. Of course, as we have seen, they do not conform to Hempel's model as such, as the explanations considered are non-deductive causal CL explanations using a default rule, rather than strict deductive arguments using exceptionless laws of nature. It would seem that something similar can be said regarding biological explanations: we have seen in chapter 7 that CL explanations are of value in biology for a number of reasons. Again, the explanations of pigeon navigation and photoperiodism are of a CL type. Moreover, unlike the explanations considered in chapter 6, these biological explanations do make use of deduction – they are DN rather than just CL explanations. Of course, the point about laws remains: although the explanations of pigeon navigation and photoperiodism are of a DN type, they too do not invoke exceptionless laws of nature, but pragmatic laws.

Nevertheless, despite these caveats, the conclusion that DN explanations do exist in biology, and moreover, are of real value in this discipline, might be termed surprising if we take a moment to evaluate the state of the debate about the relation between DN explanations and mechanistic explanations. Again, as will be evident from the quotes below, the following applies to other disciplines besides biology.

The philosophical trend is away from Hempel's model, not towards it. Among the champions of mechanistic explanations, there is a sentiment that, given its central role in the explanatory practices of biologists, the notion of mechanism has received too little attention in philosophy of science, and moreover that this neglect has something

to do with the dominance of Hempel's views on explanation, which in turn is held to be the result of a disproportionate focus on physics. Bechtel and Abrahamsen put it succinctly:

> Given the ubiquity of references to mechanism in biology, and sparseness of reference to laws, it is a curious fact that mechanistic explanation was mostly neglected in the literature of 20th century philosophy of science. This was due both to the emphasis placed on physics and to the way in which explanation in physics was construed. (Bechtel & Abrahamsen 2005 p. 423)

Although I concur with the mechanists that up until the turn of the millennium, the notion of mechanism received too little attention, I believe that this neglect has been made up for during the past decade. It is now more than ten years since Machamer, Darden and Craver's famous paper appeared, and the JSTOR website reveals that, over a 3 year period, it is the most cited paper published in Philosophy *of Science*.[68] This indicates that a lot of work has been done on mechanisms and mechanistic explanations. While I agree that mechanistic explanations are important for understanding biological capacities, I think that care should be taken not to make the same mistake Bechtel and Abrahamsen (rightly) point out in the quote above: to focus on one type of explanation at the expense of others. Given the present situation in the philosophy of biology, the warning should be that the focus on mechanistic explanations should not lead us to neglect other types of explanation of biological capacities, in particular DN explanations. In the preceding chapters I have tried to make the case that non-mechanistic DN-type explanations of capacities are *also* important in biology.

The position I defend here goes against the mainstream view that DN explanations are impossible in the life sciences or, if possible, that they are not interesting. Here are a few quotes that I think are representative of this mainstream view:

> In this sense, the covering law model is inaccurate when it states that all science consists of a search for real "general laws." One can read an entire article in *Science* on research findings in biology and not encounter anything a scientist would call a general law. (D'Andrade 1986 p. 22).

[68] *Thinking about mechanisms* was cited 32 times over the past three years (retrieved 31-05-2012 from: http://www.jstor.org/action/showMostCitedArticles?journalCode=philscie). The second best, Larry Laudan's *A Confutation of Convergent Realism* has only 20 citations; Kitcher's 'Explanatory Unification' has 11, Hempel & Oppenheim's *Studies in the Logic of Explanation* 8. *Thinking about mechanisms* is also the most accessed paper through JSTOR over the past 3 years (1316 times), but the difference with other papers is small (e.g. Hempel & Oppenheim is has been accessed 1298 times).

> No one not in the grip of the DN model would suppose that we can explain why someone hears a consonant like the speaking mouth appears to make by appeal to the McGurk effect. That just *is* the McGurk effect (Cummins 2000 p. 119).

> The received view of scientific explanation in philosophy (the deductive-nomological or D-N model) holds that to explain a phenomenon is to subsume it under a law. However, most *actual* explanations in the life sciences do not appeal to laws specified in the D-N model. (Bechtel & Abrahamsen 2005 p. 421-422)

> For these three reasons [accidental generalizations, explanatorily irrelevant premises, and the failure of nomic expectability], the CL model of explanation has generally faded from philosophical currency. Also for these three reasons, the CL model is not an especially useful starting place for thinking about the norms of explanation in neuroscience (Craver 2007 p. 40).

As is clear from these quotes, the DN and CL models are getting pretty bad philosophical press these days. Moreover, on the whole the protagonists seem to have little appreciation for the varieties within the CL model (with the possible exception of Craver 2007). Without denying the importance of mechanistic explanations, I take issue with the common view that DN and *a fortiori* CL explanations are impossible and/or uninteresting in biology. In effect, I defend a complementarity thesis: biology needs both mechanistic explanations and non-mechanistic DN explanations.

8.2 A third contender?

Having said that, I am the first to acknowledge that the brand of DN explanations we considered in the previous chapter depart from Hempel's original model in some important respects. At the close of this third part of the dissertation, let us take a moment to reflect on this.

Recall the conditions of adequacy for DN explanations, as found in Hempel and Oppenheim 1948, p. 137:

Logical conditions of adequacy for DN explanations:

*R*1: The explanandum must be a logical consequence of the explanans.
*R*2: The explanans must contain general laws, and these must be essential for the derivation of the explanandum.
*R*3: The explanans must have empirical content, that is it must be capable, at least in principle, of test by experiment or observation.

Empirical condition of adequacy for DN explanations:
*R*4: The sentences in the explanans must be true.

Obviously, there is a problem with *R*2. As I have already indicated, the status of laws in biology is controversial. How are we to understand generalizations like "All animals with a magnetic compass have the capacity to find the way back home on clouded days" (L4) and "Biloxi soybeans are short-day plants which start to produce and transport the flower stimulus if its leaves are exposed to a minimal dark period of 10 hours" (L10)? They are nomologically contingent and spatio-temporally restricted, to be sure, but to what extent is this bad news? At this point, let us recall a lesson Sandra Mitchell has taught us:

> Rather than bemoan the failure of biological generalizations to live up to the normative definition of exceptionless universality, the pragmatic approach suggests a different philosophical project. To understand the multiple relations among scientific generalizations one must first explore the parameters which make generalizations useful in grounding expectation in a variety of context (1997, p. S478).

According to Mitchell, these pragmatic aims include among others degree of accuracy and level of ontology (1997 p. S477). Regarding degree of accuracy, i.e. being attuned to specified goals of intervention, both L4 and L10 seem to serve this aim well, as both suggest further experiments. Second, level of ontology concerns generalizations about populations which describe "...structural relations between trait-groups" (Mitchell 2003 p. 125). In L4, these related traits are having a magnetic compass and the capacity to home, while L10 describes a connection between the traits of the capacity to expose leaves to sunlight and the production of flower stimulus. Thus, when we speak of DN explanations in biology, we need to understand the generalizations as being required to attain one or more of these pragmatic goals.

*R*3 must also be modified. Rather than relaxing however, the requirement should be made more stringent. Not only should the explanans have empirical content, it should make reference to an underlying mechanism, but without necessarily containing detailed information about its entities, activities or organization. Of course, it could

contain details about the mechanism, and in such cases, a mechanistic explanation might well be preferable – nevertheless, here we are interested in a situation in which such details are simply not available (yet). Rather than simply waiting for the gaps to be filled, scientists can and do construe explanations in these situations, to achieve (some of the) goals we considered in the previous section.

Thus, although the explanations we have considered in this chapter are non-mechanistic, and formally adhere to the DN model, the semantic requirements set them apart from Hempel's DN explanations, and as such, constitute a genuine third alternative. Let us call them DN* explanations. The conditions of adequacy for DN* explanations would then read as follows:

Logical conditions of adequacy for DN explanations:*

*R*1: The explanandum must be a logical consequence of the explanans.

*R*2: The explanans must contain pragmatic laws, and these must be essential for the derivation of the explanandum.

*R*3: The explanans must have empirical content, that is it must be capable, at least in principle, of test by experiment or observation. Moreover, it must make reference to a mechanism underlying the explanandum.

Empirical condition of adequacy for DN explanations:*

*R*4: The sentences in the explanans must be true.

With this in mind, we can offer a more precise diagnosis of what is wrong with the debate about biological explanations, as characterized in the previous section. The mechanists were right that Hempel did not place enough semantic requirements on DN explanations for them to be of great value within the context of biology; however, from this it does not follow that DN-like explanations, such as DN* explanations, are not a valuable tool in the explanatory toolkit of biologists. That is, the mechanists mistake was to reject all the formal conditions, because of a defunct in the semantic conditions. We have avoided this problem by defining a hybrid form, DN*, which keeps the argument format but also posits the presence of a mechanism.

Part 3

Model explanations

Chapter 9
Model explanations: a (second) introduction

9.1 Some general philosophical issues about models

To introduce the subject of model explanations, let us begin by briefly considering some more general issues surrounding the use of models in science, before zooming in on cognitive science and biology. A large part of the recent debate about models in science focusses on their *representational character*. As models are thought to represent something, three important philosophical questions arise. First, what do they represent? The answer to this question comes in two flavours: a realist one, according to which models represent (some part of) reality, and what one might call a theoretical one, according to which models represent something we ourselves constructed. On the realist side of things, a model might target anything from physical (e.g. the solar system), cultural, (the U.S. government), and biological systems (the human digestive system), to cognitive systems (numerical cognition), organizational structures (companies) and data (seismic readings). On the theoretical side, a model might represent constructs of our own making, such as theories, laws or mathematical equations.

For the purposes of this dissertation, I will take the realist stance, as I will specify below. Now a second question arises: how to characterize this representational relation between model and world? This question has been debated extensively in the literature over the past years (Bailer-Jones 2003; Contessa 2007, 2010; Frigg 2010; Giere 2004; Toon 2012). Again, there are many positions to choose from here, the main ones assuming a broadly language-based view, according to which models represent reality in the same way that language represents it. Of course, the question how words can mean something is an old and vexing one, and has its proper place in the philosophy of language. On the other hand, non-linguistic positions are also possible. These positions usually draw on the notion of (partial) isomorphism (van Fraassen 1980; Da Costa & French 2003).

Finally, there is the question about the nature of the models themselves. While some models, such as scale models of buildings, astrolabes, or the Watt Governor (employed in van Gelder 1995 as a model of the non-representational mind), are clearly physical objects, other models are fictional objects. One of the reasons for this is that they are often (sometimes deliberately) idealized, and thus contain fictions. Thus, Frigg writes:

> Frictionless planes, spherical planets, infinitely extended condenser plates, infinitely high potential wells, massless strings, populations living in isolation from their environment, animals reproducing at constant rate, perfectly rational agents, markets without transition costs, and immediate adjustments to shocks are but some objects or features that figure prominently in many model systems and yet fail to have counterparts in the real world (Frigg 2010 p. 257).

As one would expect, the jury is still out on this one. Some philosophers concur with Frigg's claim that models are, or at least contain, fictions (Godfrey-Smith 2006; Contessa 2010), while others resist it (e.g. Giere 2009).

What stance one takes on these issues should depend on the context, most importantly on the scientific discipline in question. For example, it seems fairly obvious that while some models, such as models of the circulatory system, attempt to represent real systems, models in other disciplines, for example mathematical logic, do not. Similarly, the degree of isomorphism between model and reality may vary from context to context (Teller 2001). In the present context, we are interested in model explanations of biological and cognitive capacities, some examples of which we have already met. With this context in mind, let us take a stance on these issues, in so far as this is possible.

The models discussed in this part of the dissertation are schematic representations of how certain features of a system produce some overall behaviour. Regarding the first issue, I take the realist position: the models attempt to represent the actual system responsible for the target behaviour (the explanandum), although they may approach this system with varying degrees of abstraction. Second, regarding the relation between the model and the target system, the best we can do is to say that the model describes the system. This description may take a linguistic form, it may come in the guise of mathematical equations (as we have seen in chapter 6), or it may happen through a diagram, such as a flow chart. Then again, these different types of description may be used in combination. Finally, regarding the ontology of models, the examples we shall consider are usually not physical objects, but conceptual collections of entities and activities organized in a way that allows them to perform a certain capacity (see section 5.3). Recall Cummins' definition in section 5.3: all that is needed for a model is that it gives some analytical account of how a certain function arises as the result from some hierarchy of sub-functions, or, in our terminology, to give an account how a capacity arises from the organization of sub-capacities. Again, these positions are not to be seen

as definite philosophical stances on the issues considered above; rather, they simply follow from the examples of model explanations to be found in the scientific literature of the particular disciplines we have been discussing.

9.2 Functional and mechanistic models: recapitulation and statement of the main claims of this part

I have already explained the main commonalities and differences between functional and mechanistic models in section 5.3. In this section therefore, as far as the relation between functional and mechanistic models goes, I will confine myself to providing a brief recapitulation – whenever I need to refer to some of the more detailed aspects of this relation, I will refer back to section 5.3. More importantly, I will briefly sketch some of the main tenets of current thought about functional and mechanistic explanations, especially with regards to the issue of the explanatory power of such models, and conclude by stating the main claims I will defend in the chapters to come.

In the 1970s and 1980s, discussion about the explanatory potential of models often focussed on how one can explain a capacity by means of decomposition, i.e. by dividing and subdividing complex routines in terms of ever more simple sub-routines (Cummins 1975, 1980; Dennett 1987). Briefly, on these terms, to explain a capacity is to construct an analytic account A that shows how system S realizes capacity C. Because these functional models are silent with regard to what realizes a certain (sub-)capacity, they offer the life scientist a considerable degree of conceptual flexibility to characterize various cognitive and biological capacities in terms of abstract in- and output relations. One example of such functional models that we have already encountered in section 5.3 is Levelt's model of speech production.

However, it is increasingly argued that this functional framework is inadequate when it comes to explicating model explanations in the life sciences, and that instead of offering abstract analyses of a system's behaviour in terms of a hierarchy of capacities and sub-capacities, researchers actually construct models of the underlying mechanisms (Machamer, Darden & Craver 2000; Glennan 2002; Tabery 2004; Bechtel and Abrahamsen 2005). These mechanisms include not only activities, but also entities engaging in these activities and an organization such that the collective of these entities and activities produce the overall capacity we are trying to explain. For a model to be explanatory, it is thought to be necessary to include reference to all three of these elements (Craver 2006).

The current sentiment among philosophers concerning themselves with the explanatory practices of the life sciences then is to emphasize the role of mechanistic models of at the expense of the more traditional functional models. In this part of the dissertation, I will compare these two types of models with respect to some of the properties they can have or fail to have – most notably the properties of plausibility (the degree of accuracy with which a model describes a target system), richness (the amount of detail the model possess) and performance (the ability to perform a capacity *well*, which can mean performing it faster, more reliably, with greater success rate etc.). Thus, the initial decision of a scientist to construct a model explanation in the first place is by no means the only phase in the explanatory process that is of interest to philosophers of science. On the contrary, it is after this decision has made that some of the more conceptually challenging questions arise: How do model properties such as richness, plausibility and performance relate to each other? How do they relate to the general issue of the explanatory power of models? How do they function as goals or epistemic interests behind model construction? Once the decision to develop a model explanation has been taken, a lot of interesting choices remain, and what properties one considers important in a particular model (possibly at the expense of others) is decided by pragmatic factors, as we will see.

Since functional models played such an important explanatory role in the past, it seems that many of the features of mechanistic models that are currently emphasized in the philosophical debate, in particular the supposedly necessary reference to the actual entities of a mechanism, cannot account for the explanatory potential of models in general – this will be one of the main claims defended in chapter 10. I will argue for this claim by distinguishing between the plausibility and richness of models, and reflecting on how these two features relate to each other. In chapter 11, I will compare plausibility with another feature, namely performance. Based on an assessment of models in the context of engineering and artificial intelligence, I argue that plausibility is sometimes subordinate to performance.

Chapter 10
The explanatory power of functional and mechanistic models: plausibility versus richness

10.1 Introduction[69]

This chapter is chiefly devoted to the question whence model explanations derive their explanatory power. As we have seen in the previous part of this dissertation, I am by no means against mechanistic explanation. Rather, the point is that I think the recent debate about mechanistic explanations suffers from a number of false assumptions and confusions that need to be cleared up in order to arrive at a descriptively accurate picture of explanation in cognitive science and biology. In part II, I addressed one such false assumption, namely the idea that because Hempel's model of explanation was too strict, CL explanations are of no use at all. In this chapter, I want to clear up a confusion that I think hampers progress in the debate about model explanation as it is currently conducted by the protagonists of mechanistic explanations.

Although I agree that researchers in the life sciences typically explain a phenomenon by providing a model of the responsible mechanism, I think that the recent debate on mechanistic explanations conflates two features of models that, for reasons I will establish below, are better kept distinct: their *plausibility* and their *richness of information*. By plausibility, I mean the degree of probability that a model is accurate in the existence of, and distinctions between, the various entities and activities it postulates,[70] while richness concerns the degree of detail a model provides in its description of a

[69] Parts of the following sections have been published in Gervais & Weber 2013a.

[70] Accuracy should not be understood here as exact isomorphic correspondence between the model and the mechanism, but rather as a degree of similarity: the term plausibility is partly chosen to reflect this.

mechanism's entities and activities. The conflation of these two features in the debate on mechanistic explanations is undesirable, as it has led some of the participants to view both of them as necessary for a model to have explanatory power, while richness is only required with respect to a mechanism's activities, not with respect to its entities. I argue that richness about entities, although a virtue for many other reasons, is not necessary for a model to be explanatory: there are models that say next to nothing about the details of a mechanism's entities yet still have considerable explanatory power. To put it plainly, the conflation of plausibility and richness leads one to discard as non-explanatory models that quite clearly are, which, as we will see, fits poorly with scientific practice.

Let me conclude this introduction with a brief overview. In section 10.2, by considering how mechanistic models relate to the more traditional functional models of the 1970s and -80s, I will show how the debate on mechanistic explanation confuses plausibility with richness. As a focal point, we will use Craver's continuum from how-possibly to how-actually models (Craver 2006), although the confusion applies more generally to the debate on mechanistic explanations. In section 10.3, I will briefly make the case that plausibility and richness can vary independently of each other in more than one way, so that there is at least a conceptual reason to keep them apart. Finally, sections 10.4 and 10.5 bring out the true cost of the confusion, by providing examples of models that are explanatory, but offer no details about the *entities* of the mechanism they are describing. These counterexamples are meant to move the discussion beyond mere philosophical contrivances, to the actual explanatory practice of science: in the cases under consideration, the explanation of face recognition and memory. In effect, the central claim will be: plausibility and richness concerning the target mechanism's activities is necessary for model to have explanatory power, richness concerning its entities is not. In section 10.6, I will offer a brief recapitulation of the ground we have covered in this chapter.

10.2 Phenomenal versus explanatory models

As we have seen in section 5.3, in the explanatory practices of the life sciences, mechanistic models present a clear break from the old functional models that were once used, particularly in the cognitive sciences during the eighties and nineties, to explain a system's behaviour. Although once popular, functional models are increasingly criticized for remaining silent about what (neural) entities realize a given (sub-)routine. Recall again Levelt's functional model of speech production we considered in section 5.3. Of course, this is a very abstract, rough model of how the capacity to produce

speech might be realized. Moreover, even if it is accurate in the activities it postulates, can it be really said to explain speech production if it has nothing to say about the neural entities responsible for all these activities? This is the central idea behind the 'boxology' objection.

In contrast, mechanistic explanations go beyond an abstract representation of the organization of activities, and describe the actual entities thought to realize these activities. A functional model might be useful for the purposes of prediction, mapping the input-output relation of a target system, yet for a model to have explanatory power, this is not enough, so the mechanists say. After all, using the Ptolemaic model of the heavens one can predict with some accuracy the location of certain celestial bodies in the night sky, but it does not explain why the planets move the way they do. According to this line of thought then, phenomenal accuracy is not sufficient for explanation, because mechanical accuracy is necessary (compare Glennan 2005 for a similar distinction, between phenomenal and mechanical adequacy).

This brings us to a further consideration: models vary in the degree to which they are accurate. As I have mentioned in section 7.1, Craver proposes a continuum on which a given model or mechanistic explanation can be placed, depending on its degree of mechanistic plausibility (Craver 2006; the idea was first proposed in Machamer et al. 2000). The continuum ranges from speculative sketches, where a lot of the details are left out, to ideally complete descriptions, which identify every entity and activity in the mechanism. He distinguishes three developmental stages (Craver 2006 p. 361):

1. **How-possibly models**. These models are speculative conjectures about how a capacity might be realized: specifying a set of possible parts and activities that account for the behaviour of the mechanism.

2. **How-plausibly models**. These are "...more or less consistent with the known constraints on the components, their activities, and their organization" (Craver 2006 p. 361).

3. **How-actually models**. These are ideal, complete descriptions of the model. They describe how the mechanism is composed and works in reality.

So far so good. However, does this continuum only represent the degree of plausibility? Not according to Craver himself. He seems to relate it to the idea of explanatory power, for although he does state that "...how accurately a model must represent the details of the internal workings of a mechanism will depend upon the purpose for which the model is being deployed" he adds that "If one is trying to explain the phenomenon, however, it will not do merely to describe some mechanisms that would produce the phenomenon" (Craver 2006 p.361). It seems that the closer to the ideal description-side

of the continuum a given model is placed, the more justified we are in calling a model an explanation. Why is this? The idea is that as one moves from how-possibly to how-actually models, the answers one gets allow a greater degree of control: "Deeper explanations show how the system would behave under a wider range of interventions than do phenomenal models" (Craver 2006 p. 358).

At first glance, one might be tempted to identify how-possibly models with functional explanations as I described them above. However, even the how-possibly models conjecture the existence of specified parts, although we might not have any evidence that these parts exist. This would mean that purely functional models as described above are not explanatory at all, as they are deliberately silent about the realizer of a given function or sub-function. Are functional models then like the Ptolemaic model of the heavens, merely phenomenal models that only succeed in mapping the input-output patterns of the mechanism in question? Not necessarily. A model may be silent about the parts, yet accurately represent the internal activities. In that sense it might still possess a degree of mechanical accuracy.[71] The difference between descriptions of entities versus activities will become important later on.

For the moment however, let us take note of an ambiguity in Craver's continuum. He seems to be talking about two things. On the one hand, he clearly talks about the accuracy of a model; i.e. the degree of correctness or truthfulness of a model. The phrases 'how-plausibly' and 'how-actually' make this plain. Yet there is also another sense in which the continuum is ordered: from abstract descriptions, to more detailed, complete descriptions. This is evident when he talks of a how-actually models as being a "...ideally *complete* description..." (Craver 2006 p. 360 my italics), or when he says: "Between sketches and complete descriptions lies a continuum of mechanism schemata that abstract away to a greater or lesser extent from the gory details..." (Craver 2006 p. 360). In effect, *Craver conflates plausibility with richness, and holds the presence of both to be necessary for a model to be of explanatory power.*

This cluster of ideas is already present when, in their 2000 article, Machamer et al. distinguish between what they call 'mechanism sketches' and 'mechanism schemata', where the former are abstract, incomplete versions of the latter, in that their "...entities and activities cannot (yet) be supplied..." (Machamer et al. 2000 p. 18). In other words, they are less rich in details about parts and operations, and only when such information has been added are they considered to be full-fledged mechanism schemata that are required for mechanistic explanation.

---

[71] Again, compare Glennan's notion of mechanical adequacy of a model, which is, among other things determined by the degree to which it provides "...quantitatively accurate descriptions of the interactions and activities of each component..." (Glennan 2005 p. 457).

Nor are these authors alone in asserting that such a conflated notion of plausibility and richness is required for explanation. Glennan, for example, writes:

> The requirements for a model being a description of a mechanism place substantive constraints on the choice of state variables (such as the fact that state variables should refer to properties of parts), parameters and laws of succession and coexistence. The satisfaction of these (...) constraints is what accounts for the explanatory power of mechanical models (2005 p. 448).

while Bechtel claims that advancing a mechanistic explanation

> ...requires decomposing the mechanism into component parts and operations and localizing each operation in the appropriate part (Bechtel 2007 p. 176).

To recapitulate, it is reasonable to distinguish between mere phenomenal and genuinely explanatory models. In the literature on mechanistic models, there is a tendency to claim that what is required for a model to be of explanatory value, is that it specifies the actual mechanism responsible for the explanandum. However, we have seen that this is really a two-part job: the model has to provide a description of the parts and operations of the mechanism, and it has to do so in an accurate, informed way. That is, the model should be rich in that it gives details about entities and activities, and these details should be plausible.

10.3 Plausibility and richness vary independently

However, at least conceptually, there are good reasons to keep these two features a model can exhibit separate. A higher degree of plausibility, for example, does not entail a higher degree of richness. It just means there is more evidence that the model is a true description. Likewise, models may be rich in detail and yet be plainly wrong. For one, though incorrect, the Ptolemaic model of the heavens is very rich in detail, postulating deferents and epicycles, fixed stars, nine celestial spheres etc. In a sense, the more details one gives, the greater the chances of error, and the more abstract a model is, the more likely it is to be right in the few assertions it does make (to the point of triviality). Let us hold on to Craver's terminology for plausibility, and introduce the terms how-abstractly, how-partially and how-concretely to cover richness. We now find ourselves with two continuums.

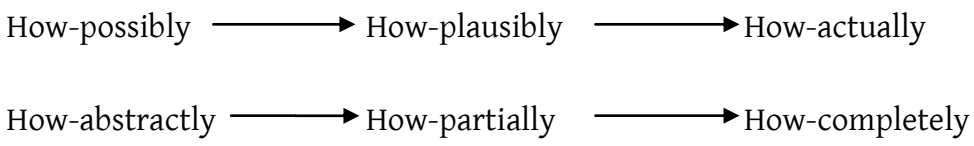

Figure 2 Two features models can have ordered on two continuums

In principle, every combination is possible. There might be complete models that are only very loosely supported by evidence (how-completely/how-possibly), highly abstract models that nevertheless enjoy considerable support in the few assertions they do make (how-abstractly/how-actually), etc.

There is an additional complication however. Recall that mechanisms comprise both entities as well as activities. Both plausibility and richness, in all their respective strengths, can apply independently to entities and to their operations. This means that the conceptual space of possibilities we have so far explored is extended with yet another dimension. A model might contain a lot of information about a mechanism's parts while little about the operations they perform. Yet the little it says about the operations might be far more plausible than the detailed story it gives about the entities. Then again, in another model the degrees of plausibility and richness might more or less converge with respect to activities, while strongly diverging when it comes to the entities.

Anyway, it should be clear that at least conceptually, plausibility and richness can vary independently from one another, both with respect to entire mechanism descriptions and with respect to particular aspects of mechanism descriptions. But of course, this is only part of the story: there remains the issue of explanatory power. By holding both plausibility and richness with respect to entities as well as activities, as necessary for explanation, one cannot make sense of the explanatory power of functional models described in the previous section. It is here that the cost of conflating these two notions is truly felt, as it hampers our understanding of the explanatory practices of the actual scientists themselves. In the next sections, I will bring this point home by considering two functional models: one of face recognition (section 10.4) and one of memory (section 10.5).

10.4 Plausibility and richness in models of face recognition

Face recognition, or the capacity to spot faces from among other sensory data is a socially advantageous trait, as it enables us to make judgments about fitness, sex, health and emotional status of other individuals. Indeed, this capacity is so highly developed

and common in humans, that it often argued to have a genetic basis (Wilmer et al. 2010; Zhu et al. 2010). In fact, we are so attuned to faces that sometimes the ability is triggered by certain features of non-face objects (e.g. distance and size ratios of rocky protrusions on a mountain surface). The idea that face recognition is a special case, i.e. requires its own explanation, separate from the general models for object recognition, is based primarily on evidence that the inversion of faces affects the ability to recognize faces more than it does other objects: the so-called *face inversion effect* or FIE (Yin, 1969).[72] Early explanations of the capacity of face recognition relied heavily on functional analysis (e.g. Marr and Nishihara 1978, Rhodes 1985). Let us here consider the example of Bruce and Young's functional model of face recognition (Bruce and Young 1986).

They explained human face recognition in terms of a functional model (see figure 3), in which visual data of a presented face is structurally encoded to produce two different types of descriptions (view-centered and expression-independent), which in turn are analysed in three different ways (analysis of facial speech and expression apply to the former, face recognition units or FRU's analyse the latter). Choosing to remain silent about the details of the first two types of analysis, they ascribed to each face recognition unit a number of stored structural codes, each one describing a face that had already been seen before and was stored in memory. The activation of these face recognition units signals the appropriate 'personal identity node' or PIN to become active. The PIN allows access to semantic information about the individual in question: name, occupation, age, and any further information about the person available. This information is stored within the more general cognitive system. Thus, name generation, i.e. the ability to put names to faces, is a serial process involving name retrieval from data stored in PIN's by FRU's.

[72] This so-called face-specificity hypothesis is not without its critics however. For a survey of the evidence and the literature regarding this issue, see Kanwisher & Yovel (2006).

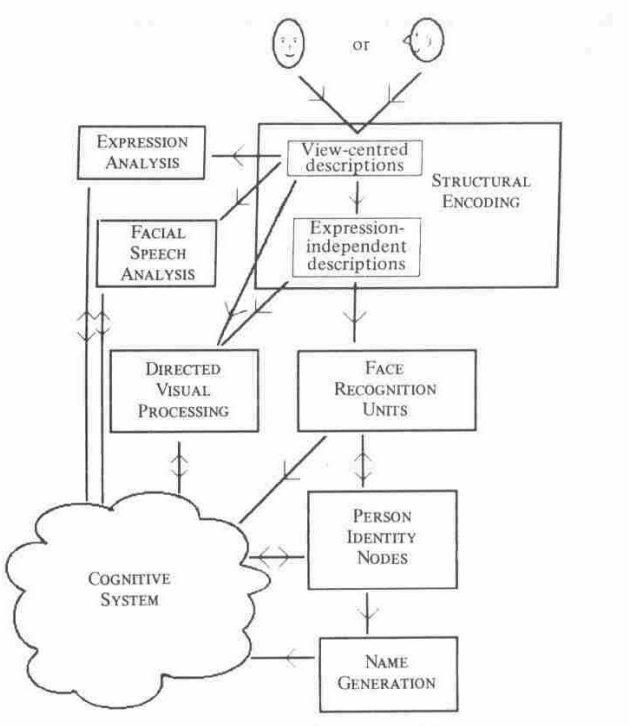

Figure 3 Functional model of face recognition (source: Bruce and Young 1986)

Here we have an example of a classic functional explanation. Bruce and Young said about their model: "...we are concerned almost exclusively with evidence in favor of functional components in the human face processing system, without regard to whether or not these are localized to specific areas of the brain" (Bruce and Young 1986 p. 306). In fact, they acknowledged the threat of boxology: "We recognize that the differences in the statuses of the arrows and the boxes used in models of this type are problematic" (Idem p. 311). However, speculative as this analysis may be, it is not simply fantasizing. There are reasons why the boundaries between the sub-functions have been drawn the way they are: "A 'box' represents any processing module, or store, which plays a distinct functional role, and whose operation can be eliminated, isolated or independently manipulated through experiment or as a consequence of brain damage" (Idem p. 311). For example, the reason to distinguish between sub-tasks having to do with person recognition (the 'person identity nodes' box) and with face recognition (the 'face recognition units' box) is that in experiments, it was found that face recognition can break down while person recognition remains intact (Hécaen 1981).[73] Moreover, although generally not interested in the neural localization of the functions they proposed, Bruce and Young did use evidence obtained from experiments on people

[73] This type of independency between two cognitive capacities (where one can break down while the other remains intact and vice versa) is called double dissociation. Another example of double dissociation will be discussed in section 10.5.

whose face recognition skills were impaired (prosopagnosia) to support their model (Idem p. 315). In fact, it seems that this functional model suggests ways to intervene upon the causal process for the benefit of experiment: just like Craver requires explanatory models to do, Bruce and Young's model shows "...how the system would behave under a wider range of interventions than do phenomenal models" (Craver 2006 p. 358).

Moreover, although Bruce and Young's model remains silent about the entities of the mechanism responsible for face recognition in humans, its fundamental functional layout is still regarded as basically correct, albeit as a crude approximation. The past decades have seen the publication of a substantial amount of research papers on human face recognition, complete with new evidence from experimental psychology and cognitive neuropsychology. Although alternative models featuring different cognitive architectures have been proposed (e.g. Burton, Bruce & Johnston 1990; Brédart et al. 1995), the evidence itself seems does not favour these alternatives over Bruce and Young's original model (see Hanley 2011 for an overview of the literature). In fact, when it comes to the face naming stages distinguished by Bruce and Young, it has become possible to locate plausibly corresponding anatomical locations in the brain. For example, there is evidence that the FRU's Bruce and Young described are located in the fusiform face area (Haxby, Hoffman & Gobbini 2000; Rothstein et al. 2005), while recent fMRI studies have suggested the right anterior temporal lobes as the locus where semantic information about people is stored (Tsukiura et al. 2010). Of course, this is still far from showing that Bruce and Young's model is an ideally complete description of the responsible mechanism, but it does show how functional models, which focus only on the activities of the mechanism, might later be augmented to include information about the responsible entities as well.

Returning to Bruce and Young's original model, where does it fit on the continuums we considered in the previous section? With regard to plausibility, the researchers themselves admit it is speculative. However, as we have seen, part of the model was based on experimental evidence, so we can count this model as at least plausible to some degree, which in any case is all that is needed. The first condition is thus met. What about richness? The model is rich in information regarding the mechanism's activities (analysing, processing, encoding generating), yet it says next to nothing about what realizes these activities. In fact, *it is a how-abstractly model with regard to entities.* Here we can see the true cost of the confusion we mentioned in the first section: if we do not distinguish between plausibility and richness, and within the latter, between richness regarding entities and regarding activities, then we are forced to say that this model does not explain human face recognition. However, it seems that by leaving out

the details by which all the sub-routines are implemented, the model in fact highlights just those features of the mechanism that are relevant to explaining face recognition across individuals.[74] Of course, to philosophers of science, this need not come as a surprise. As Strevens has pointed out, far from being obstructive, abstraction can actually boost explanatory power: "The salient but irrelevant causal details are the shallows, then, and the more abstract – that is, more general – properties of the system are its depths, fleshed out by the details..." (Strevens 2008 p. 137).[75]

To conclude, this model is not like the Ptolemaic model of the heavens: it postulates operations and makes distinctions based on experimental evidence, and so is more than a mere input-output mimicking device. It provides information about what goes on between input and output, and does this in a plausible way. However, it does not provide information on the parts or entities of the mechanism, and in that sense, does not meet the condition of richness. If we were to follow Craver, we would have to discard Bruce and Young's model as 'merely phenomenal' and lacking explanatory power.

10.5 Plausibility and richness in models of memory

As the example of Bruce and Young's model of face recognition indicates, plausibility and richness, while independently variable, can exert mutual influence. Frequently, considerations of plausibility in the form of experimental evidence constrain the richness of a model: add any more information, and the risk of getting it wrong increases. Again, I hold that while at least some plausibility is necessary for a model to be explanatory, richness is only required with respect to a mechanism's activities. To appreciate this, consider a model which combines jet extremer positions on the two

[74] In fact, in technological contexts, we often want our models to be rich in details about activities, rather than entities. When it comes to artificial systems, we demand of our models that they duplicate the function of a natural system, not that they accurately describe the way in which they are implemented; this is the subject of the next chapter.

[75] Even so, I should be clear where my position differs from Strevens's. According to Strevens, explanations can have depth along what he calls the "physical axis" and the "abstract axis" (2008 p. 136), where these two do not run contrariwise. Although the abstract axis may presumably be equated with the how-abstractly, how-completely continuum, the physical axis may not be so equated with the continuum of plausibility. I am talking about abstraction versus descriptive accuracy, while Strevens is talking about abstraction versus relations of causal influence (i.e. the degree to which an explanation taps into the most fundamental explanatory level).

continuums depicted in figure 2 than Bruce and Young's model: a functional model of memory.

Of course the term 'memory' is highly unspecific, it is a (folk-)psychological label that groups together a whole range of capacities, capacities which, in certain contexts, should be distinguished. Nevertheless, the fact that such diverse activities can be grouped together under the right level of abstraction only illustrates the flexibility of functional models. What level of abstraction is relevant has to do with contextual information – the only assumption here is that there is *a* context in which it makes sense to look at memory as a whole.

In any case, this treatment of memory as a single capacity is short-lived once we engage in functional analysis, which is of course what we need to do if we want to construct a model. It is common to decompose the capacity to memorize information into three sub-processes: *registering*, *storing* and *retrieving*. Of course, from a cognitive standpoint, the activity of storing is of particular interest. Typically, this activity is decomposed into two sub-routines: short-term and long-term memory (STM and LTM). Thus, we get a very rough picture, functionally equivalent to Levelt's analysis of speech production:

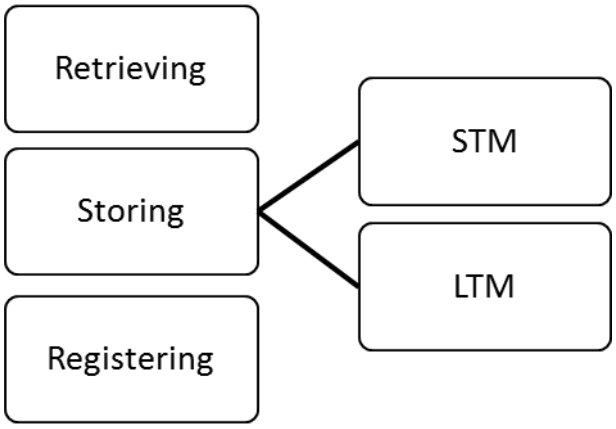

Figure 4 Functional model of memory

However vague this sketch might be, notice that like Bruce and Young's model, the boundaries between the sub-activities are not drawn at random: they are drawn to accommodate experimental evidence. In the case of the distinction between STM and LTM, the evidence is very well known. First, it was found that there is double dissociation between specific types of information storing: some brain-damaged patients are found to be impaired on their ability to memorize repeated sequences of words or numbers, while performing normally in experiments testing spatial location, retrieving read information and face recognition (Shallice and Warrington 1980); while other patients exhibit the precise opposite pattern (Milner 1966). Second, the distinction mirrors that between the so-called *recency* and *primacy* effects. When test

subjects have to memorize lists of words, if the process of recalling those words starts soon after the end of the presentation (a few seconds), they tend to be better at recalling to lasts words on the list (recency effect). If the delay is increased by a few seconds, this effect disappears. The primacy effect is the tendency of test-subjects, when the rate of presentation is decreased (there is more time in between the presentations), to better recall the *first* items on the list. Generally, these effects seem inversely related to each other, which indicates that there are two different storage systems active in memory.

But of course, LTM and STM can be analysed still further by making subsequent divisions. For example, neuropsychological evidence has suggested that LTM should be subdivided into implicit and explicit memory systems, or between an episodic and semantic memory (Tulving 1972). Similar subdivisions can be made for STM.

All these divisions and subdivisions need not detain us here. What matters is that on the basis of empirical evidence like the sort provided by double dissociation experiments, we can have proof that an activity within a mechanism operates in a certain manner without knowing any further details of either the operation itself or its parts. A functional model of memory like this has considerable explanatory power, in that it accounts for a range of explananda of which the ones mentioned above are a comparatively small but representative subset. Of course, some might object that this is little wonder, since these experimental data have been used to constrain the model in the first place. As with Bruce and Young's model, the complaint might be that this model is guilty of boxology: if the contents of the boxes are left unspecified, this type of analysis does little more than re-describe the explanandum. This would be unfair, for two reasons.

First, the objection seems to hinge on the *order of discovery and model construction*. If one first postulates a model and subsequently finds out it agrees with (new) evidence, then it can be said to explain that evidence; if on the other hand a model is constructed to fit the (already available evidence), then this is seen as vacuous. In practice however, the distinction between these two sequences of events will never hold as rigidly as presented here: typically, a model will be constructed within the room left by the available evidence, while incorporating new evidence as it appears. Second, the explanatory power of such functional models is further exemplified by the fact that the richness it has with respect to its operations has implications about its parts. In the case of memory, possible anatomical locations for the episodic LTM have been suggested (Baddeley 1998). In fact, what seems to be the case with Bruce and Young's model is repeated here: after the functional analyses, which combines the available evidence with an account of the (relations between) the operations, implementational richness follows suit.

Again, what this example makes plain is that although considerations of plausibility and richness often exert mutual influence, they nonetheless vary independently. Regarding its position on the two continuums we have described in section 10.3, it will

be clear that it is a quite rough sketch (although, as we have seen, many more subdivisions can be made, and in fact have been made), so that with regard to richness of activities, the model is still more abstract than that of Bruce and Young: clearly, it is how-abstractly. As far as plausibility goes, we have seen that the model is constructed to account for much experimental evidence, so that while perhaps not attaining the how-actually status (as this is more a regulative ideal) it does exhibit a high degree of plausibility. In fact, by joining two positions that are situated on still more extreme sides of the continuums than Bruce and Young's model, the functional model makes an even stronger case that plausibility and richness can vary, although they often exert mutual influence. What it shares with Bruce and Young's model however, is that it remains silent about the neural realizers involved and yet, *pace* Craver and the mechanists, has considerable explanatory power.

10.6 Conclusion

In this chapter, I have argued for two claims: first, that the literature on mechanistic explanations tends to confuse two features of models, plausibility and richness, and second, that this confusion can cloud our view on the practice of scientific explanation because both features are deemed necessary for a model to have explanatory power, while in fact, richness is only necessary with respect to a mechanism's activities. Regarding the first claim, I have argued that plausibility and richness come in degrees that can vary independently from each other. Regarding the second, I have argued that traditional functional models constitute counterexamples, case in point being the functional models of face recognition and memory. These models go beyond merely input-output mapping to provide information on the operations of a mechanism that secures their status as explanations, yet they are neutral about the entities that perform the operations they stipulate. As such, they underline the conclusion that while plausibility is required for explanatory power, richness is so only with respect to the activities, and not the entities, of a mechanism. For these reasons, plausibility and richness are best kept distinct.

Chapter 11
Plausibility and performance in model explanations

11.1 Introduction

In the previous chapter, we have seen that plausibility and richness can vary independently, and that consequently, we cannot declare one or other as the most important feature of models generally, that is, regardless of context. As is evident in the cases of face recognition and memory, the fact that a model lacks richness does not principally rule out any explanatory potential. Nevertheless, although in a given context (for example, when a capacity needs to be explained of which the underlying mechanism is largely unknown), one may, for a variety of reasons (including availability of evidence, predictive power etc.), make due with less details, or with less plausibility, it might still seem that plausibility and richness are generally speaking *desirable properties* of a model. According to this school of thought, while it is true that there are situations in which we must weigh the one against the other, nevertheless, ideally our models of the mechanism underlying a certain capacity should exhibit both properties to the highest degree.

However, there are two contexts in which this maxim does not hold true. These are: a) generalization and b) artificial intelligence. Regarding generalization, it is obvious that if we want out model to apply to more than one individual system, or even across species, we have to generalize, making allowances for differences between individuals. Here, richness necessarily suffers, though plausibility need not.

Given this situation, one might still maintain that while richness is not always desirable, plausibility is, though sometimes subordinate to other interests. Yet even this is not necessarily true. In the engineered models of artificial intelligence, plausibility is sometimes wholly sacrificed in the name of yet another interest: *performance*. Briefly, the idea is that while biological or cognitive plausibility may be a good starting point

when designing systems that are to accomplish certain tasks, when it comes to the punch, plausibility may in fact hamper performance, and is therefore sometimes jettisoned altogether. This is what can happen when we study artificial, as opposed to biological systems. It is for this reason that I have chosen to speak of the more neutral 'system', rather than organism.

There are two questions at stake here: 'How models generalize?' and 'What properties do we look for in models of artificial systems?' As will be made clear below, these issues are interrelated, and considering them together offers some new ways in which to expand our explanatory toolbox. In section 11.2, I will distinguish several how-questions to serve as a framework in which models of artificial systems can be discussed – the latter task will be taken up in section 11.3. Finally, in section 11.4, I will offer some concluding remarks.

11.2 Degrees of generalization and biological constraints[76]

As I have mentioned in section 5.3 (note 49), when dealing with *capacities*, it is sometimes more appropriate to say that explanations are answers to how-questions. Again, I will discuss the issue of how-questions in more detail in chapter 13, but for now, let us simply note that it has been argued in the literature that how-questions are valid explanation-seeking questions in their own right (Scriven 1962; Salmon 1989). Briefly, while answers to why-questions typically consist of identifying or referring to causes, answers to how-questions can take the form of models. In the case of the traditional functional models described in section 5.3, the basic format looks something like this: if we want to explain a capacity *C* of a system *S*, we have to construct a functional model *M* showing how *C* is performed, such that for each input, output and input-output relation in *S* there is a corresponding input, output and input-output relation in *M*. That is, if we want to answer a question like:

1) How is *C* realized in *S*?

we might construct a model *M* that maps the input-output relations that make up *C*. Having done that, we can answer 1) by saying:

[76] The remainder of this chapter has, with some adaptations, been published in Gervais (in print).

2) *C* is realized in *S* the same way that *C* is realized in *M*.

Note that although it looks like 2) just restates the mystery, it does not, for we must remember that *M* is not a mechanism or system in nature, but a model that we have constructed ourselves, so that we know in detail how it realizes *C*. However, the question seems to ask something beyond mere input-output mapping. Returning to the example of face recognition, consider:

3) How is the capacity to recognize faces realized in the human brain?

As it stands, some face-recognition systems have been developed that perform this capacity very well, in that they are able, in experimental setups, to map the input-output relations of the brain (they are presented with examples of faces and non-faces and are able to tell the difference with more or less the same degree of accuracy as humans), but do so in a fundamentally different way. Up until recently for example, they could only use two-dimensional geometrical data. Of course we do not want to count:

4) The capacity to recognize faces is realized in the human brain by applying algorithms to exclusively 2-D geometrical data.

as an answer to 3). As we know ourselves to see e.g. chins and noses as protrusions, 4) is clearly implausible. Beyond this appeal to 'first person knowledge' however, there is also some 'harder' evidence. For example, 2-D face systems notoriously suffer from what is known as the 'lighting problem': their ability to recognize faces deteriorates significantly when the strength of the light coming from the image they are presented with is varied, while humans tend to retain their abilities in such circumstances (Adini, Moses & Ullman 1997). No matter how perfectly such systems may (otherwise) mimic our performance in this task, we have to concede that, being 2-D, they are not explanatory models for face recognition as it is performed by humans.

Granted, a model may to a certain extent map the human input-output relation for a capacity, without being explanatory with respect to the human realization of that capacity. However, do models always have to be models of a capacity as it is performed in a specific (set of) system(s)? The pragmatic approach we have used so far says that if a capacity is the explanandum, the explanans can be viewed as an answer to a how-question. There is nothing to restrict this type of question to include only capacities *as they are realized in some system*, we can also ask how-questions about capacities *as such*, that is, without any particular biological or neurological constraints. That is, instead of 1), we might ask:

5) How is *C* (as such) possible?[77]

The point here is not that researchers will actually be interested in how capacities could be realized without *any* constraints: capacities are of course always realized in some system. Rather, the point is that one can have legitimate motives in placing *as little constraints on the system as possible*. In the next section, I will consider one context in which this strategy is commonplace, namely the context of engineering. For now, note that at least in psychology and the cognitive sciences, asking explanatory questions about capacities *as such* forms an important part of scientific practice, if only as a preliminary strategy (that is, preliminary to the business of answering the question how the capacity is realized in some particular system). In fact, this was already noted by Dennett back in 1978:

> Faced with the practical impossibility of answering the empirical questions of psychology by brute inspection (how *in fact* does the nervous system accomplish *X* or *Y* or *Z*), psychologists ask themselves an easier preliminary question: How could any system (...) possibly accomplish *X*? This question is easier because it is 'less empirical'; it is an engineering question, a quest for a solution (*any* solution) rather than a discovery. (...) Seeking an answer to such a question can sometimes lead to the discovery of general constraints on all solutions (...), and therein lies the value of this style of aprioristic theorizing. (...). For instance, one can ask how any neuronal network with such-and-such physical features could possibly accomplish human color discriminations (...). Or, one can ask, with Kant, how anything at all could possibly experience or know anything at all. Pure epistemology, thus viewed (...) is simply the limiting case of the psychologist's quest (Dennett 1978, 110-111).

Thus viewed, the 'Kantian' question (How is X possible at all?) can be interpreted as constituting the extreme end of yet another continuum, while enquiries about how a particular system performs that function occupies the opposite end:

[77] Note that this question does not fall into the category of Craver's how-possibly questions (2006). For Craver, how-possibly questions are loose inquiries that are made in the early stages of an investigation, in which a lot of data is still missing: they are attempts to put some initial constraints on the explanandum, prior to constructing a more informed (how-plausibly), and ultimately ideally complete description (how-actually). Nevertheless, how-possibly questions in Craver's sense are still asked with respect to a capacity as it is performed by some system. The question under consideration differs because it is asked about a capacity *as such*, regardless of any particular realization.

How does *S* perform *C*? How is *C* performed in *S* and $S^{1...n}$? How is *C* possible at all?

Figure 5 Different levels of abstraction at which one might seek to explain a capacity

As Dennett notes, it is possible to begin with more general questions, discovering constraints having to do more with *C* itself than any particular realization of it, and work your way to a particular realization of *C* in *S*. However, explanation can also work in the opposite direction. As one moves to the right of the spectrum, the number of constraints will decrease. This means that somewhere along the line you get to the point where *C* is described in such a general way that it applies to more than one system. In other words, the scope increases. Examples of this can be found in medicine. Suppose that an impaired capacity in a brain damaged patient has somehow been restored by the brain. In this case, we might be interested to know just exactly how that capacity is carried out in this damaged brain. Then again, if we have a satisfactory answer to this question, we may want to devise a model of how that capacity is realized that is sufficiently general to apply both to the patient and healthy human beings. In circumstances like these, we are actually looking to move from the left hand side of the spectrum toward the middle. Of course, detail matters: as soon as we reach the point where all the relevant systems fall under the scope of that capacity, we stop. In the example just given, as soon as we have described the capacity in such general terms that it applies both to healthy patients and the brain damaged patient, we stop jettisoning constraints, as plausibility still matters enough not to be sacrificed without just cause. This stopping has to do with our methodological interests: it is simply the act of eliminating variables.[78]

In the next section, I will give a more detailed example of this explanatory strategy. For now, note again that richness or level of abstraction is a matter of degree. How many constraints one places on the system responsible for a certain capacity will be decided by pragmatic issues. This however, seems at odds with Craver's continuum as described in section 10.2, which endorses descriptive accuracy about implementational details as necessary for a model to have explanatory power. Of course, this is particularly striking for questions located near the right end of the spectrum: surely, one cannot expect a model answering 5) to excel in descriptive accuracy, for there is no mechanism specified to describe. In fact, *any* model of *any* system that realizes *C* is a valid answer. Presumably, scientists are rarely (if ever) interested in capacities under no constraint whatsoever. The continuum sketched in figure 5 is related to the richness

[78] Also, think of animal testing: here we continue to drop constraints until the capacity is described in such a way as to apply across species. Again, *S* can be any system, natural or artificial.

continuum depicted in figure 2, but it is reversed to indicate that explanatory progress is not always a case of filling in more details – sometimes, it involves exactly the opposite: if what we want is an increase in scope, we want abstraction – though again, plausibility matters in that we stop generalizing as soon as we have reached the desired level of generality. This is because implicit in the present discussion is still the idea that models represent how a capacity is realized by (sets of) *natural* systems. In the next section however, we shall discuss models of *artificial* systems.

11.3 Model explanations in artificial intelligence

It might seem that our discussion about explanation-seeking how-questions about capacities as such is somewhat contrived. Despite Dennett's remarks about the value of discovering general constraints that are inherent in capacities as such, one may wonder whether the kind of Kantian question he refers to has any bearing on scientific practice at all. To see that it does, I will now turn to modelling in engineering contexts. After all, one branch of cognitive science is artificial intelligence, and in this discipline, type 5) questions do arise. Here, the explicit goal is to design and construct artificial systems that can perform the same cognitive capacities we associate with human intelligence. Let us consider one specific example of a cognitive capacity: exact calculation.

Humans are endowed with the capacity to perform exact calculations accurately, up to a certain level of complexity. If we ask how we perform this capacity, the model that answers this question indeed derives its explanatory power from (among other things) its neurophysiological plausibility. That is, suppose we want to answer:

6) How is the capacity to perform exact calculations realized in humans?

The model that we use to answer this question has to reproduce the capacity under a number of constraints. For example, some artificial computing devices might make poor models, as they diverge from human brains in important respects: they might be neurophysiologically implausible, or they might fail to reproduce the capacity to perform exact calculations (e.g. they might be less exact, or they might take far longer to solve arithmetic problems).

However, although these respects are important to contexts like the one referred to in question 6), there are other contexts in which they are less important, or downright irrelevant, and these other context might still have to do with explaining the capacity. In other words, descriptive accuracy or plausibility is not the only explanatory context in which we could be interested in the capacity: there are other reasons we might want

to explain the capacity to perform exact calculations. Suppose an engineer wants to construct a desk calculator. Of course, his goal is not to construct a model of how humans perform complex calculations: after all, he is designing a tool that, hopefully, surpasses our own ability. In fact, he seeks to *duplicate* the capacity. Motivated by this interest of duplication, he might ask:

7) How is exact calculation as such possible?

This may sound somewhat artificial. In fact, when constructing a desk calculator, there are all kinds of constraints he needs to take into account.[79] The point is that these constraints are different from the ones applying to exact calculations as it is performed by humans. Thus, a sensible strategy would be to put fewer constraints on the capacity, until the scope is broad enough apply to both humans and certain artificial devices. In terms of the continuum sketched above, we stop somewhere in the middle, at the point where the scope is just broad enough to encompass both the human realization of the capacity and an artificial one. To put it in other terms, we stop where the forces pulling in opposite directions, namely level of detail (to the left) and duplication (to the right), balance out for the task at hand.

But that is not all. In some engineering contexts arising in the field of artificial intelligence, it is not uncommon to jettison the requirement of plausibility completely. To appreciate this, let us continue to pursue the example of the engineer trying to construct his desk calculator. There are a number of models that can perform exact calculations. For reasons of clarity and briefness, let us consider classic computationalism and connectionism. The symbolic architecture of classic computationalism, where symbols are manipulated according to a pre-programmed set of rules, is very good at performing very complex calculations with great accuracy, far surpassing that of any human. On the other hand, as a model of the mind, computationalism is outdated. The serial nature of its operations and its consequent brittleness does not compare to the robustness of our brains. Connectionism on the other hand, resembles our brains more closely. In fact, in the original debate between computationalism and connectionism as candidate models for the mind, the latter's neural plausibility (in the form of distribution of activity over a network of nodes, graceful degradation, its ability to recognize patterns etc.) counted as an important

[79] Examples of such constraints are: the materials available, convenience of use and time considerations (we want the calculator to perform calculations rapidly – within a timeframe that is of use to us, that is).

point in its favour (McClelland and Rumelhart 1986).[80] However, despite all these advantages, they perform poorly when it comes to exact calculations. In fact, connectionist networks have in the past been ridiculed for answering a question like "What is two plus two?", after much crunching, with "About four" (Boden 2006 p. 964).

Clearly, exactness is a virtue when it comes to desk calculators. In fact, when engineering interests drive model construction, *performance trumps plausibility*.[81] Duplication therefore, is only a subsidiary goal: it is really the desire to make a system that outperforms humans that motivates the engineer, and the model he finally constructs will reflect this. Of this model, and of the flow chart representing how the calculator performs the exact calculations, we can say three things. First, with regard to how humans perform exact calculations, it is an inaccurate model and fails to explain it. Second, with regard to how the calculator performs it, it is an ideally complete description and explains it, but that is hardly surprising, since it is the very blueprint the engineer used to make the calculator in the first place. Third, with regard to the capacity to perform *exact calculations* as such, it explains how that capacity *can be* performed. When the engineer asked 7) and started decomposing exact calculation down into sub-routines, he *was* looking for an explanation, only not with neurophysiological plausibility on his mind, but performance.

Yet there are other interests besides duplication or performance that might prompt the search for an explanation of such capacities. Another interest is *unification*. Once an artificial system has been designed and constructed, then to anyone besides the engineers involved in this process of designing and construction, the explanatory question might arise as to what these artificial systems have in common with e.g. natural systems. Thus, one might ask the following question:

8) How is the capacity to perform exact calculations performed in this desk calculator and in humans?

This question is situated somewhere in the middle of the continuum presented in figure 5. In effect, what we are asking for here is what two realizations of the capacity of exact calculations have in common with each other. Again, I will consider the issue of how-questions in more detail in chapter 13, but for now, note that these comparative

[80] As the debate currently stands though, connectionist networks are considered to be highly idealized models too – but still more plausible than classic computationalist architectures.

[81] To appreciate the complex relationships between plausibility, duplication and performance, consider this remark: "Although the biologically-inspired models are very useful for neuroscientists, ultimately, when building a commercial face recognition system, one should use the algorithm with the highest performance, regardless of biological relevance. However, for specialized applications, (...) models developed using human psychophysical evidence might outperform other algorithms" (Sukthankar 2000).

question-types are often motivated by generalization: in revealing features that are common to the operations of both types of systems, an answer to 8) brings together information from multiple and diverse sources. And of course, an answer to comparative question-types like 8) will typically take the form of a model – a functional model. In the case of question 8), this is especially clear, since any similarity between humans performing complex calculations and desk calculators exercising the same capacity will not be found in the entities, but will be confined solely to the domain of the operations. Yet, despite its abstract nature, such a model would clearly be of explanatory value to those who are interested in the similarities between human and artificial performances of exact calculation.

Thus, we have seen that the interest of plausibility can be subordinate to the interest of performance and the (secondary) interest of duplication. Doubtless, all this does not tarnish the explanatory importance of mechanistic models when it comes to explaining capacities as they are realized in particular systems. Of course we need the models of e.g. biological functions to be accurate, and not only phenomenally adequate. It might follow that for particular systems, this accuracy is necessary for a model to have any explanatory power regarding that capacity. What does not follow however, is that phenomenal and functional models have no explanatory power in *any* context. Reiterating Dennett's point, asking about capacities under fewer constraints can be a valuable research strategy. Ultimately, how many constraints one takes into account is decided by one's interests: in the case of performance, an interest typical of engineering contexts, these constraints will surely be determined by practical considerations having to do with performance, rather than not empirical considerations having to do with plausibility. Nevertheless, this does not undermine the explanatory power of answers to such questions. However, although strictly speaking correct, this conclusion should not be the main point to take away from this discussion, if only for the fact that Craver and the mechanists have a very different context in mind from some of the one considered here. Of greater importance is the observation, borne out by the continuum sketched in the previous section and illustrated in this section, that the business of explaining capacities by constructing models is far more diverse and dynamic than the literature on mechanistic explanations suggests.

11.4 Some concluding remarks

Two final remarks are in order. First, although distinct, engineering and plausibility interests are often present at the same time and can act complementary. This is

especially the case when a model has to be constructed of a capacity at which, unlike exact calculations, humans are particularly good. Face recognition for example, is a capacity in which we excel, and many of the early artificial systems badly underperformed compared to us, being sensitive to all kind of distortions (we already encountered the lighting problem, faces presented at angles is another one) that human test persons just see right through. In such cases of course, an engineer wanting to design such an artificial system has everything to gain by first asking how the capacity is realized in us. The point is though, that even here, plausibility is only an intermediary goal. As soon as artificial systems are starting to equal or outperform us, engineers will drop plausibility as a goal, as it no longer serves the greater goal of performance.[82]

Finally, one may wonder whether the capacities targeted by functional explanations in engineering contexts, such as the one described in the previous section, are still properly called *cognitive* capacities. Can we still talk of subtraction as a cognitive capacity when it is performed by a humble desk calculator instead of a person? Here, one might point out that the engineering sciences (artificial intelligence in particular) have a history of fruitful interaction with the cognitive sciences. Artificial systems can help us understand our own capacities, while knowledge of these may in turn lead engineers to improve the performance of these systems. After all, the point made in this chapter is that plausibility is sometimes sacrificed entirely for the sake of performance, not that it *always* sacrificed.

[82] And in fact, with the example of face recognition systems we considered earlier, this is beginning to happen right now; see the results from the 2006 Face Recognition Vendor Test (available for download at: http://www.frvt.org/).

Part 4

From explanation to explanation-seeking questions

Chapter 12
The significance of why-questions[83]

12.1 Introduction

In the previous two parts of this dissertation, I have applied the pragmatic approach to various debates about explanation in the cognitive sciences and biology. This part will be more theoretical in nature. Here, I will reflect on the motivation (i.e. the epistemic interests) behind, and the relations between, different question-types. In this chapter, I will focus on how one can use the epistemic interests to answer questions about the significance of explanation-seeking questions and their answers; that is, on what makes explanation-seeking questions, and the explanations themselves, interesting for both scientists and philosophers (thus avoiding the green cheese and red herring problems discussed in chapter 3). This chapter focusses on why-questions; in chapter 13, I will consider how-questions.

In chapter 3, we have seen that van Fraassen's pragmatic or erotetic theory of explanation faced two severe challenges, namely the green cheese and red herring problems. To reiterate, the objections ran that van Fraassen fails to put sufficient restrictions both on what counts as a legitimate why-question and on what counts as a relevant answer to a why-question. The critics point out that the erotetic model faces two types of counterexamples: (1) there are why-questions which we intuitively judge as not legitimate but are legitimate why-questions according to the erotetic model (the green cheese problem); and (2) there are answers that we intuitively see as inadequate explanations but are adequate according to the erotetic model (the red herring problem).

[83] This chapter is based on Weber, Gervais & Van Bouwel 2013.

This chapter has four aims:

1) I want to show that *each* of the two problems leads to *two* tasks that philosophers of science can set for themselves: a *theoretical* task and a *practical* task. The theoretical tasks are about fundamental epistemological insights: they require us to formulate necessary conditions for explanation-seeking questions and their answers to be significant. The practical tasks are of a methodological nature: they require us to formulate heuristic guidelines (they are about helping scientists to ask the right why-questions and giving adequate explanations). So there are *four* tasks for philosophers of science originating from the two problems.

2) I want to show that the two theoretical tasks are very difficult: their accomplishment requires that we satisfy two desiderata that pull in opposite directions.

3) I want to show that the practical tasks are easier to handle, because there is no such tension and they allow for a piecemeal approach.

4) I want to make some substantial contributions to the two practical tasks.

Of course, since one of the main goals of this dissertation is to argue for robust pluralism instead of van Fraassen's anything goes pluralism, this last aim is of particular importance.

The structure of this chapter is as follows. In section 12.2 I show how the green cheese and red herring problems lead to the theoretical and practical tasks (cf. my first aim). In Section 12.3 I show that the theoretical tasks are very difficult (cf. the second aim) and in this way motivate my pessimism about them (I think that the chance that someone will accomplish them is very low). Next, in Section 12.4 I deal with the practical task that originates from the green cheese problem, and in Section 12.5 I will address the practical task that originates from the red herring problem. So these two sections taken together deal with the third and fourth aim. Finally, in section 12.6 I will put the issues covered by this chapter in the context of the somewhat broader recent discussion of contextualism. I will do this by comparing the results with two contextualist solutions to the green cheese and red herring problems that have been proposed in the literature: Risjord's laissez-faire contextualism (Risjord 2000a) and Khalifa's hands-on contextualism (Khalifa 2004). Of particular importance is the comparison with Khalifa's position, since his project fits into the practical tasks I will delineate.

Although the examples in this dissertation are generally taken from cognitive science and biology, in this chapter, I will focus more on the social sciences, as the recent

discussion about the green cheese and red herring problems (in Khalifa 2004 but also in chapter 4 of Risjord 2000a) also happens to be situated in that context.

12.2 Two problems, four tasks

To refresh our memory, in 3.2 we have summarized van Fraassen's view on what counts as an appropriate explanation-seeking question, and what counts as an appropriate answer to such a question, as follows:

> **α)** It is worthwhile to attempt to answer the contrastive question "Why X rather than Y?" if and only if (a) X is true and (b) Y is false.
> **β)** An answer to a contrastive why-question is an adequate explanation if and only if (a) the question is about a true contrast and (b) the answer is true and stands in the contextually determined relevance relation R to X.

As we have seen in section 3.3 however, this leads to problems, as van Fraassen's erotetic model of explanation suffers from excessive liberalism, both with respect to what counts as an appropriate question, and what counts as a relevant answer. What can philosophers of science do to remedy this situation? That is, what can philosophers of science do to bolster anything goes pluralism into a more attractive position?

The *theoretical* tasks which philosophers of science can set themselves in reaction to this situation are in fact attempts to do better than van Fraassen. The first possible theoretical task is to fill in scheme α in a better way than van Fraassen did:

> **α*)** It is worthwhile to attempt to answer the contrastive question "Why X rather than Y?" if and only if [...] .

The green cheese problem means that van Fraassen's way of filling in the "[...]" in the scheme is not adequate. So we can try to do better. If we manage to accomplish this first task in an adequate way, we obtain an important epistemological insight: we know what makes a contrastive question a real explanation-seeking question.

Analogously, the red herring problem leads to the possible theoretical task of giving a better implementation of scheme **β**:

> **β***) An answer to a contrastive why-question is an adequate explanation if and only if [...]

If we manage to accomplish this second theoretical task, we obtain another important epistemological insight, viz. what makes an answer to a contrastive question an explanation.

The *practical* tasks which philosophers of science can set themselves are (1) try to formulate guidelines which help scientists to avoid uninteresting ("green-cheese-like") questions and (2) try to formulate guidelines which help scientists to avoid uninteresting ("red-herring-like") answers. While the theoretical tasks are of a descriptive nature, the practical tasks are methodological: the aim is to offer *strategies* which help scientists to avoid asking uninteresting questions and considering irrelevant answers to be adequate explanations.

12.3 The difficulty of the theoretical tasks

In this section I want to show that the two theoretical tasks are very difficult: their accomplishment requires that we satisfy two desiderata that pull in opposite directions. Since the difficulties reveal themselves very clearly when you try to fulfil the task, this is what I will do: first we try to do it as good as possible (12.3.1 and 12.3.2) and then will I show that the attempt is not really satisfactory and point out that there is a tension between opposite desiderata (12.3.3).

12.3.1 Sophisticated pragmatism

The attempt at solving the theoretical tasks starts from a view on the aims of sciences which we might call *sophisticated pragmatism.* The main idea is that the aim of science is not just to provide a true description of the world. In *Science, Truth and Democracy*, Philip Kitcher formulates an argument supporting this view:

> Nobody should be beguiled by the idea that the aim of inquiry is merely to discover truth, for, as numerous philosophers have recognized, there are vast numbers of true statements it would be utterly pointless to ascertain. The sciences are surely directed at finding significant truths. But what exactly are these? (2001: 65)

Kitcher's answer is double. On the one hand, there is *practical significance*:

> One possible answer makes significance explicitly relative – the significant truths for a person are just those the knowledge of which would increase the chance she would attain her practical goals. Or you could try to avoid relativization by focusing on truths that would be pertinent to anyone's projects – the significant truths are those the knowledge of which would increase anyone's chance of attaining practical goals. (2001: 65)

But for Kitcher there is more:

> Neither of these is at all plausible as a full account of scientific significance, and the deficiency isn't just a result of the fact that both are obviously rough and preliminary. Linking significance to practical projects ignores areas of inquiry in which the results have little bearing on everyday concerns, fields like cosmology and paleontology. Moreover, even truths that do facilitate practical projects often derive significance from a different quarter. Surely the principles of thermodynamics would be worth knowing whether or not they helped us to build pumps and engines (and thereby attain further goals). Besides the notion of practical significance, captured perhaps in a preliminary way by the rough definitions given above, we need a conception of "theoretical" or "epistemic" significance that will mark out those truths the knowledge of which is intrinsically valuable[84]. (2001: 65)

Because there is something more than just practical significance, I will call a view on the aims of sciences like Kitcher's and mine *sophisticated pragmatism* as opposed to the view – which could be labelled *strict pragmatism* – that science should only look for practically significant truths (this view was defended by John Dewey).

Let us now apply this view to contrastive explanations. The general idea of sophisticated pragmatism is that the aim of science is to find *significant truths*. Applied to the explaining of contrasts we get: the aim is to find *significant true* answers to why-questions about *significant true* contrasts. If we agree that this is the aim that scientists must have when they are explaining contrasts, this imposes restrictions on what scientists should do.

First, this aim implies a condition of significance for contrastive questions:

> **α**)** It is worthwhile to attempt to answer the contrastive question "Why X rather than Y?" if and only if (a) X is true and (b*) the contrast is significant.

Let us compare this to what van Fraassen said:

[84] Note that the term "intrinsically valuable" is ill-chosen. What Kitcher means is: valuable for non-practical reasons.

α) It is worthwhile to attempt to answer the contrastive question "Why X rather than Y?" if and only if (a) X is true and (b) Y is false.

The difference is clear: my clause (b*) versus his (b). As I have already mentioned, I believe that the requirement that Y is false is clearly too strong and should be dropped. Moreover, while van Fraassen's theory is unable to avoid the green cheese problem, clause (b*) of our condition of significance of why-questions (which says that it is only worthwhile to answer a contrastive question if the contrast is significant) allows to cope with it. Of course, this does require us to say something about the notion of 'significance' (otherwise, we would only solve the problem by stipulation). I will do so in section 12.3.2.

The aim of finding significant true answers to why-questions about significant true contrasts also implies a restriction on answers:

β)** An answer to a contrastive why-question is an adequate explanation if and only if (a*) the question is about a significant true contrast and (b*) the answer is significant and true.

Now recall the corresponding claim of van Fraassen:

β) An answer to a contrastive why-question is an adequate explanation if and only if (a) the question is about a true contrast and (b) the answer is true and stands in the contextually determined relevance relation R to X.

Again the difference between van Fraassen and my view concerns my reference to the concept of significance, both in (a*) an (b*). Where van Fraassen's view on what counts as an adequate explanation provides no answer to the red herring problem, the extra clause of significance in my proposal does. This again raises the question: what is significance?

12.3.2 Significance

To avoid excessive liberalism, my approach relies on the notion of significance, both with respect to why-questions and their answers. How to spell out this notion? I think there are two common ways in which a topic and foil may significantly relate.

First, it may be that the topic is contrasted with a foil that is *at odds* with the topic. These situations are common in science. To give an historical example, consider the

question: "Why does the orbit of Uranus deviate from the orbit predicted by the 1845 standard model of our solar system, while the orbits of all the other planets are as predicted by this model?" With "1845 standard model" we mean Newton's laws plus the assumption that there are 7 planets (from Mercury up to Uranus, but without Neptune). As is well known, Neptune was discovered in 1846 and its gravitational effect on Uranus explains the deviations. Borrowing some terminology from Nozick (1981), let us call these foils which make the topic rather unexpected *apparent excluders*.[85] Apparent excluders do not operate in isolation: they make the topic unexpected if we combine them with (often implicit) background knowledge. For instance, in the Uranus case there is the assumption that there are no undiscovered planets.

Second, it may be that the foil represents a more *desirable situation* than the topic: it represents an improvement. Again, such situations occur frequently in science. For example, it is the case that the hole in the ozone layer over the Antarctic is expanding; since it not expanding is the more desirable situation, one might reasonably ask: "Why does the hole in the ozone layer over the Antarctic increase in size, rather than decrease or remain stable?" Such questions are asked with an eye to bring about a situation that is an improvement over the current situation.

The foregoing considerations give us enough material to construe a tentative[86] definition of significance with respect to why-questions:

> *Significance of contrastive questions*
> The contrast between topic X and foil Y is significant if and only if Y relates to X in one of the following ways:
> (a) Y is an apparent excluder of X, or
> (b) Y represents an improvement over X.

[85] Nozick's format is: "How is *X* possible, given that *E*?" where *X* is the newly observed data, and *E* is the theory or body of evidence that seems to exclude *X*. This format is useful not only to understand certain explanation-seeking questions in science, but also in other, more exalted contexts. In theology for instance, we might ask: "How is evil possible, given that there exists an omnipotent, omniscient and benevolent God?" Similarly, philosophers of mind have asked: "How is mental causation possible, given the causal closure of the physical domain?" An answer to these questions should alleviate the intellectual tension brought about by the inconsistency, either by rejecting *X* (because of some equipment malfunction, measurement fault or whatever) or, by rejecting *E* (for example in a Kuhnian-style revolution).

[86] This definition is tentative because of certain shortcomings I will point out in 12.3.3. However, I will not present a better one, since I do not have a better one – in that sense, the definition is not tentative: it is simply my best shot.

As we have seen in the previous subsection, Kitcher distinguishes between epistemic and practical significance. Option (b) is an implementation of the idea of practical significance for contrastive questions. Option (a) is an implementation of the idea of epistemic significance for contrastive questions.[87] A brief example: the question "Why is the level of unemployment in Belgium 7.4 %, rather than the moon made of green cheese?" is not significant according to our definition.

A similar tentative definition of significance can be developed for answers:

> *Significance of answers to contrastive questions*
> A statement A is a significant answer to a contrastive question Q if and only if
> (a) A dissolves an apparent exclusion in Q, or
> (b) A suggests a means of attaining a foil that is an improvement over the topic.

Again, the options (a) and (b) are implementations – for the case of answers to contrastive questions – of the ideas of epistemic and practical significance. A brief example: "Because red herrings have gills" is not a good answer to the (significant) question "Why is the level of unemployment in Belgium 7.4 %, rather than 10.7 %?" because it does not do any of the things specified in our definition of significance of answers.

Before we discuss their shortcomings, let me point out that these two definitions certainly do not constitute an exhaustive theory of significance, for two reasons. First, they only deal with why-questions and explanations, while there are numerous other scientific activities for which one can try to define significance. Second, within the domain of why-questions and explanations, we have only considered contrastive questions: significance of non-contrastive why-questions is beyond the scope of this chapter. In other words: the scope of these definitions is limited, so they do not constitute an exhaustive theory of significance.

12.3.3 Evaluating the definitions

Now that I have done my best to accomplish the theoretical tasks, let us see how well I did. An important characteristic of the two definitions is that they are *disjunctive:* they tell us that there is significance if at least one of two conditions is fulfilled, but they do not tell us what the two possibilities have *in common*. In other words, they do not give

[87] There are other possible implementations of these ideas. I will come back to this in section 12.3.3. This is the most important reason why the theoretical tasks are so difficult.

insight into what significance *is*. They just list two possible cases. In logic and mathematics this is known as "definition by cases". Definition by cases may be a good instrument for representing our intuitions (if they lead to correct appraisals, see below) but they do not provide much philosophical insight. Even if the definitions do not suffer from counterexamples, they are not what we really want: because of their disjunctive nature, they do not tell us what the "nature" of explanation-seeking questions or the "nature" of explanation is.

A second problem with the definitions is that they are too restrictive: in order to avoid false negatives (intuitively significant questions and answers that are nevertheless excluded by the definitions) we have to add more disjuncts. To see this, let us adapt a question we used in the previous subsection. Suppose the hole in the ozone layer remains stable in year *x*. Then we can ask:

> Why does the hole in the ozone layer over the Antarctic remain stable this year rather than increase in size?

This question is similar to the one in 12.3.2, but topic and foil are switched. The underlying motivation is not improvement, because the foil is not the ideal state. A possible reason for asking this question is *preservation*: we want to preserve what is good. This is a practical motivation for asking why-questions which is complementary to the one I have included in the definition (improvement). In order to avoid false negatives, we have to take such questions into account by adding a third possibility. The definition of significant answers must also be adapted.

The two problems together result in a tension. In order to have a definition which is philosophically interesting, we need to find some overarching principle so that we can eliminate the disjuncts. In order to account for our intuitions, we have to add more disjuncts. And more disjuncts make it more difficult to find an overarching principle that covers them all. Because of this tension, the theoretical task is extremely difficult. If someone accomplishes it, we will finally know what makes a contrastive question a real explanation-seeking one, and what makes an answer an explanation. However, because of the tension just explained, I am pessimistic.

12.4 Significant contrastive questions

As we have seen at the end of section 12.2, there are two practical tasks which philosophers of science can set themselves: formulate guidelines which help scientists to avoid uninteresting ("green-cheese-like") questions; and formulate guidelines which

help scientists to avoid uninteresting ("red-herring-like") answers. In this section we develop and discuss guidelines of the first kind. The second type will be given in Section 12.5.

Two points should be noted about these guidelines. Unlike the necessary conditions for significance that the theoretical tasks demanded from us, these guidelines give sufficient conditions. They are heuristic tools to help scientists select significant questions and answers. As such (and this is the second point), they are future-oriented.

The guidelines for contrastive questions conform to the following pattern:

> If conditions c_1, ..., c_n, hold then question X is a significant explanation-seeking question.

So the guidelines are positive: they are rules of thumb for formulating significant contrastive questions. In the next two subsections, I will consider four such guidelines. I certainly do not claim that these guidelines are jointly exhaustive. One of the reasons why I do not claim exhaustiveness is that I focus on contrastive questions about particular facts. Similar guidelines can be formulated for other types of why-questions. And even if we restrict ourselves to contrastive why-questions, many more positive guidelines could be developed.

12.4.1 I- and I'-type questions

Here is the first guideline conforming to the pattern described above:

> Suppose that object *x* has property P at time *t.* Then the question "Why does *x* have property P, rather than the ideal property P*?" is a significant explanation-seeking question.

Where P and P* are mutually exclusive properties. We call questions of the type considered in this guideline *I-type questions* because invoke an ideal state. Before I give an example and justify this guideline, let me present a second one which is closely related:

> Suppose that object *x* has property P at time *t* and object *y* (which like x belongs to class C, or bears relation R to x) has the ideal property P* at *t.* Then the question "Why does *x* have property P, while *y* has the ideal property P*?" is a significant explanation-seeking question.

Let us call the questions referred to in this guideline *I'-type questions* because (like the I-type) they invoke an ideal state but in a different way (see section 12.5.1).

Let us consider an example of each type. Suppose there are two fields of potatoes (*a* and *b*) which are regularly infected with late blight (*phytophtora infestans*, the oomcyete or microorganism that caused the 1845 potato famine in Ireland). Now about one of these fields, a researcher might pose an I-type question:

Why has field *a* been struck by late blight, rather than remain healthy?

This type of question contrasts an observed state of affairs with an ideal one, i.e. one we consider to be preferable. According to our first guideline, it is an interesting explanation-seeking question. Suppose now that suddenly, during harvest time, it is found that the crops of field *b* have remained healthy, while field *a* is struck by late blight. In this case, researchers might pose an I'-type question:

Why has field *a* been struck by late blight, while field *b* remained healthy?[88]

This question contrasts two observed state of affairs, one of which is the ideal one. The two objects have something in common: they both are crop fields. According to the second guideline, this is a significant explanation-seeking question.

Let us now turn to the justification of the guidelines. The two guidelines I have presented here follow from the fact that improvement is an epistemic interest. As I have said in section 4.2, epistemic interests are types of motivations for scientists to search for explanations, but they also function as motivations for other people such as policy makers and the general public, to be interested in the explanations scientists give and to pay them for their research. Improvement in the sense of making a given situation better is certainly one of these typical motivations, and as we have seen in the case of the Challenger disaster in chapter 4, so are prevention and responsibility. Finally, the interest of preservation we encountered in section 12.3.3 can play a similar role, acting as a motive not just for scientists to provide explanations, but also for laymen and policy makers to be interested in them.

[88] Thus, here we have an example of a question that contrasts two state of affairs that are both true. As I accept these questions as scientifically valid, I side with Lipton (1990 p. 250) in rejecting van Fraassen's third claim as listed at the end of section 3.2, namely that all foils are false.

12.4.2 E- and T-type questions

Here is a third guideline:

> Suppose that object *x* has property P at time *t* and objects y_1, ... y_n (which like *x* belong to class C or bear relation R to *x*) have property P*. Then the question "Why does *x* have property P, while objects y_1,..., y_n have property P*?" is a significant explanation-seeking question.

Let us call questions of the format used here "E-type" questions. The fourth guideline is:

> Suppose that object *x* has property P at time *t* and property P* at time *t'*, and we have a reason to expect no evolution in *x* in this period. Then the question "Why does *x* have property P* at *t'*, while it had property P at *t*?" is a significant explanation-seeking question.

Let us call questions of the format used here T-type questions. E- and T-type questions have foils that apparently exclude or diminish the chance of the topic occurring. Let us now consider an example. Suppose it is observed that a western European country *c* has no increase in unemployment while its neighbours have. Then we might pose an E-type question:

> Why did the unemployment figure of *c* remain stable given that it rose in the surrounding countries?

Given the increase in the surrounding countries, we had expected the figure to rise in *c* as well. That is, the foil is an apparent excluder of the topic, and hence, the question is significant. The answer to this question then should tell us where our expectations went wrong.

Suppose that it is further observed that this rise in unemployment in the surrounding countries has happened quite suddenly. We might then ask a T-type question (*d* is one of the neighbours):

> Why has the unemployment figure of *d* risen during the last year, while it remained unaltered the year before?

This question contrasts a current state of affairs with a past one that is not the case anymore. In this example it is an unexpected evolution in time that is the target for explanation.

Note that one could also ask a question about the *non*-occurrence of a certain evolution. In the example under consideration, if the unemployment figure has not risen, while we had reasons to believe it would (for example if there was a monetary crisis going on), then it is the *absence* of change which is in need of explanation. Let us call that a T'-type questions. In both cases (T and T') it is a surprise *given some preconceived opinions or arguments* that prompts the question. The transition between the two periods can be marked by change or continuity, but is always unexpected. Similarly, we can define E'-type questions as the complement of E-type questions: E'-type questions are about unexpected similarities between objects or events.

The justification of these guidelines again relies on epistemic interests. In the previous subsection the epistemic interests were of a practical nature. Here we have to invoke more theoretical ones, such as *resolving apparent exclusions* or more generally, the desire to unify our knowledge and make it more coherent (resolving apparent excluders is a specific way to increase coherence).

12.5 Significant answers

Let is now turn to guidelines for answers. We will show that it is possible to formulate guidelines of the following form:

> If the question is a significant contrastive explanation-seeking question, and the motivation for raising it is of type X, then an answer that satisfies condition Y is a significant answer.

In such guidelines "condition Y" is the relevance criterion that is supposed to ensure that the scientist who follows the guideline gives a significant answer. As we will see, Y can be complex (made up of several sub-conditions). Note that this format refers to types of motivations for asking why-questions (epistemic interests). Their role will become clear in what follows.

Many guidelines of this form can be developed. In section 12.5.1 I will propose a guideline for I- and I'-type questions. In Section 12.5.2 I will present one for T-type questions. As with the guidelines for significant questions, I do not claim exhaustiveness.

12.5.1 Answering I- and I'-type questions

According to an article in Time Magazine of January 1, 1945, the US army, fighting on the European continent in the last wet months of the previous year, suffered an outbreak of *foot immersion syndrome*, also known as trench foot (after its devastating impact at the western front during the First World War). Trench foot is a condition brought on by poor vascular supply to the feet, due to prolonged exposure to damp and cold conditions. Symptoms include numbness, swelling and early stages of necrosis in affected areas. According to the article, approximately 17,500 U.S. soldiers developed the condition. An I-type question would be:

> Why did the American soldiers suffer from trench foot, rather than stay healthy?

The epistemic interest which motivates this question might be *improvement*: we might be interested in keeping the soldiers healthy, and discovering the cause of trench foot is of course instrumental in doing so.[89] This last point is important: in posing this type of question with a therapeutic motivation, the researcher determines that – in order to be significant – an answer to his particular question must have *causal relevance* to the contrast in that question. In the case at hand, causal factors in developing trench foot are prolonged exposure of feet to cold and damp conditions; conditions that are typical of northwest Europe during November and December.

But there is more. The article goes on to say that in contrast, the British army, even though it made its way through the damp plains of Holland, reported no similar problems, as its soldiers wore robust gum boots, which they were required to keep waxed, and were instructed to regularly massage their feet and change their socks. In general, the British were ordered to take care of their feet, while the Americans were not. This information makes it possible to pose an I'-type question:

> Why did the American soldiers suffer from trench foot, while the British soldiers did not?

The motivation behind this I'-question can be improvement, but the added bonus over the I-type question is that the answer you get to the former typically suggests that there is a solution which is not beyond our reach.

[89] Of course, this is not to say that the journalist who actually wrote the article had any interest beyond simply reporting an outbreak of trench foot (although, being an American himself, he would in all likelihood not have been entirely neutral). However, army officials may very well have this interest.

The explanation might look like this:

> The American army experienced an outbreak of trench foot because:
> 1) the conditions were damp and cold; and
> 2) the soldiers did not take care of their feet.
>
> The British army did not experience an outbreak of trench foot even though:
> 1) the conditions were damp and cold, because
> 2) the soldiers took care of their feet.

I will use this example to illustrate the first guideline for answers:

> If an I-type or I'-type question is asked, and the motivation for asking it is "improvement", then an answer that (a) gives causal information, (b) highlights causes that make a difference and (c) at least partially cites causes that relate to possible human interventions, is a significant answer.

This guideline contains three conditions. Let us look their justification. That the answer has to give causal information follows immediately from "improvement" interest which is in the antecedent part of the rule. Condition (b) is necessary because not all causal information is relevant, as Peter Lipton succinctly argues:

> Suppose that my car is belching thick, black smoke. Wishing to correct the situation, I naturally ask why it is happening. Now imagine that God (or perhaps an evil genius) presents me with a full Deductive-Nomological explanation of the smoke. This may not be much help. The problem is that many of the causes of the smoke are also causes of the car's normal operation. Were I to eliminate one of these, I might only succeed in making the engine inoperable. (1993, p.53)

Finally, the condition that at least one of the causes must relate to a possible human intervention (like the soldiers taking or not taking care of their feet) directly follows from the "improvement" motivation: if there is no such cause, we cannot ameliorate the situation.

12.5.2 Answering T-type questions

Next, I will consider significant answers to T-type questions. Recall that the foil is, by definition, always an apparent excluder of the topic. So there is some surprise and unexpectedness involved.

I will discuss answers to T-type questions with an example of the social sciences: the explanation of the *gender gap.* The term *gender gap* refers to differences between men and women on various public and private issues in the political and social sphere. Specifically, social scientists have been interested in how these gender related differences play out in the respective voting behaviour of men and women. During the 1950s and 1960s, men and women in the US exhibited very similar voting behaviour, as both tended to favour right-wing parties (there was no significant gender gap); however, from the 1980s on, this pattern started to change, as women were reported to offer disproportionate support to left-wing parties compared to men (Chaney, Alvarez & Nagler, 1998; Manza & Brooks 1998).[90] Thus, the gender gap occurred, constituted by a shift in voting behaviour of American women over time as compared to men. About this phenomenon, researchers asked a T-type why-question of the form:

> Why did American women offer more support to left-wing parties in the 1980s and 1990s, while on average they voted more conservative in the 1950s and 1960s?

Why did researchers ask this question? It was motivated by surprise: they had reasons to believe that no change would occur. In fact, early post-war sociology did not focus on gender at all (Stouffer 1955; Lipset 1960), the general view among social scientists being that women simply lacked interest in politics (Berelson et al. 1954, p. 25). It was thought that when women did cast their vote, their acceptance of the traditional family roles simply made them follow their husband's choice (Campbell et al. 1960). Given these preconceived opinions, the T-type question they asked was motivated by surprise: an unexpected evolution in time had to be explained. Therefore, a significant answer to a question like this should tell us why things are different from how we expected them to be.

The answers researchers provided cited causes that were located at the level of U.S. politics. That is, social scientists tried to explain this shift in political preferences by referring to factors specific to the U.S. political situation, such as party differences on the ERA (a proposed amendment to the American constitution guaranteeing equal rights for women under federal, state and local law) and strong divisions on topics as abortion and welfare reform policies (Costain & Berggren 1998; Mueller 1988).

This example illustrates the following guideline:

---

[90] Thus, the gender gap refers specifically to a gender based divergence in voting behaviour, not to the way this divergence is actually situated on the political spectrum: if women had continued to vote conservative while men changed to liberal, this would also have constituted a gender gap. Of course, the explanation would be quite different in such a case.

> If a T-type question is asked, then an answer that (a) gives causal information, (b) highlights causes that are present in the first period and absent in the second (or vice versa) is a significant answer.

The justification for this guideline is that the two conditions ensure that the apparent exclusion in the question is dissolved.

Let us explore this example somewhat further. The explanations scientists gave for the gender gap in the U.S. leads to further expectations: insofar as the causes of the gender gap are specific to the political situation in the U.S., we would not expect to find a similar pattern in other countries, where these factors are absent.

However, Giger argued that evidence obtained from the EuroBarometer suggests that in most Western-European countries, a similar shift in political preference among women is discernible, though appearing somewhat later and in varying degrees across countries (Giger, 2009). Thus, Giger was led to ask a T-type question:

> Why does voting behaviour among women of Western European countries change, rather than remain unaltered?

This question was also motivated by surprise: given the fact that the gender gap was previously explained by referring to conditions specific of the U.S., the result was unexpected for everyone who accepted that explanation. A significant answer to this question should ideally indicate what led us to the wrong expectation: it should refer to factors that were previously ignored. Giger argues that the developmental theory of gender realignment (henceforth DTGR) by Inglehart and Norris (2000, 2003) does a better job of explaining the phenomenon of the gender gap, to the extent that it also explains its occurrence in Western Europe. Rather than pointing at specific circumstances of any one country, DTGR points at structural and cultural developments that are common to wealthy, post-industrial societies. These include: reforms in the paid labour force for women, more equal opportunities of education and a shift in traditional family values. By referring to these factors, the theory explains why the gender gap also manifests itself in Western European countries and the developed world in general (rather than post-communist or third world societies), and reveals what features were left out the original explanation and why this omission led us to wrong expectations.[91]

[91] Which is not to say that the U.S. specific explanation contradicts the one provided by Giger: for all we know, the U.S. specific factors are still explanatorily relevant for explaining the gender gap in the U.S. In that sense,

12.6 Two contextualist solutions

Finally, let us take a moment to put the issues covered in this chapter in a somewhat broader philosophical perspective, by comparing my proposal to two contextual solutions to the green cheese and red herring problems that have been offered in the literature. As the name suggest, these solutions mean to rule out problems of excessive liberalism by regarding the context in which a particular why-question is asked. However, contextualism comes in different flavours.

First, let us consider Risjord's *laissez-faire contextualism* (2000a). Basically, this approach tries to avoid the problems by shifting the focus from analysing why-questions in *general*, as Kitcher and Salmon have done (Kitcher & Salmon, 1987), to the study of the different (epistemic interests) involved in constructing *specific* why-questions. The idea is that in scientific practice, the general philosophical worries of excessive liberalism simply never apply. Thus, concerning the red herring problem, Risjord simply says: "The laissez-faire contextualist response to the red herring problem is that *in a given context*, it is not the case that anything can explain anything" (Risjord, 2000a p. 82). About the green cheese problem: "As far as the presuppositions go "the moon is made of green cheese" is a foil that could be appropriate in any context [...] Again, the burden of precluding such absurd foils falls to the interests of the investigators" (Idem, p. 82). Thus, in practice, scientists are deemed mature enough not to consider green cheeses as foils, nor red herrings as appropriate answers to the why-questions they raise. The relevance criterion is restricted by their interests. As Risjord acknowledges, this puts a substantial burden on these interests, but he goes on to argue that this is no reason for hard-core philosophers of science to worry, since interests are not always the 'passing fancies or whims' they are typically held to be: they "...are shared, and not accidentally, by groups of people. Shared environment, experience (including education), and common problems all contribute to the interests of a group" (Idem, p. 83). Thus, in another study Risjord puts this view into practice, showing how non-epistemic (in this case political and moral) interests have historically influenced the choice of contrast class and relevance criterion in anthropology (Risjord, 2000b).

This seems to solve the problem from a practical point of view. However, I do not think that as far as scientific practice is concerned, the green cheese and red herring problems have ever been considered very urgent problems to begin with. It is only

---

the first answer, though it led us to entertain wrong expectations, was significant in its own right. Rather, Giger's explanation is better in that it highlights different causal factors that are *also true* of the U.S. That is to say, these causal factors might have been used to explain the gender gap in the U.S., and if they had been, we would not have been led to entertain false expectations (namely that the gender gap is specific to the U.S.).

philosophers who are bothered by them, and from a philosophical point of view, Risjord's laissez-faire contextualist solution seems indeed aptly named. Rightly, I think, Khalifa warns that this approach may lead philosophers to simply report what scientists in the field are doing, without any ambition to construct an overall account of why they are doing what they are doing (Khalifa, 2004 p. 40). To wet the philosophical beak, a more robust defence of the erotetic model is needed. Of course, this is not to say that we should resort back to trying to accomplish the theoretical tasks as I described them in section 12.2, but rather that we should take the practical task, i.e. the task of formulating practical guidelines, seriously. In this sense, Khalifa's hands-on contextualism seems a more promising approach. So what does Khalifa's approach look like?

Khalifa (2004) shares the opinion that van Fraassen's theory of explanation is too thin to exclude the kind of ridiculous questions and answers exemplified by the green cheese and red herring problems. Let us first check what his aim is and how it relates to the theoretical and practical tasks we have distinguished. According to Khalifa, the challenge that results from the green cheese problem is "to offer principles for narrowing down the contrast class" (2004, p. 45). This is reminiscent of the first practical task we have distinguished. Khalifa's response to the challenge consists in three *strategies* (his term) which can be used to arrive at good questions (see below). This is clearly an attempt to fulfil the practical task we have described. Khalifa's paper does not contain a definition of significant questions, so he does not attempt to achieve the first theoretical task. That is also the case for the answers to the questions: Khalifa gives guidelines, but his paper does not contain a definition of what an explanation is. So he does not want to fulfil the second theoretical task either.

Khalifa presents three strategies for identifying foils. First, the topic and foil tend to refer to the same thing: the so-called *semantic strategy*. Second, the topic and foil are confirmable by the same data-generating procedure; let us call this the *methodological strategy*. Third, the most relevant foils will be those that are the expected results or implications of a plausible hypothesis; we might label this the *epistemic strategy*. For Khalifa, these three strategies are complementary: together they identify relevant foils, and it is their complementary application that allows scientists to avoid "green-cheese-like" questions.

Let us now compare the strategies to what we have done in Section 12.4. We begin with the epistemic strategy: the idea that the most relevant foils are the expected results or implications of a plausible hypothesis. How does this work? Khalifa presents a detailed (and convincing) case study on how the procedures of measuring occupational status and social mobility used in an influential 1960s study by social scientists Peter Blau and Otis Dudley Duncan, shaped the contrasts in their why-questions and the way they assessed the explanatory merits of their answers. Khalifa writes:

> Blau and Duncan provisionally began with the hypothesis that African American social mobility was analogous to white social mobility but revised it as the data bore out certain differences. Whites coming from the lowest socioeconomic origins had the highest chance of upward mobility. From this Blau and Duncan conjectured that African Americans should, on average, be more likely to experience upward social mobility than whites since a larger portion of the African American community was concentrated in these lower socioeconomic strata. However, the data confirmed quite the opposite [...] From this, we get the following why-question: Why do African Americans have a lesser (rather than greater) chance of upward mobility than white Americans?" (2004, p. 46).

One can easily see the similarities between this situation and the one I described in the explanation of the gender gap: given a certain situation, we come to harbour certain expectations, and when these expectations are disconfirmed, this leads us to pose a contrastive why-question that is motivated by surprise. As Khalifa says: "...expectations play a crucial role in identifying which propositions will be included in the contrast class" (2004, p. 47). Obviously, this epistemic strategy for choosing foils constitutes a similarity between our proposal and Khalifa's hands-on contextualism: the third and fourth guideline, as described previously correspond to this idea.

However, there are two important differences. The first is that I have templates for the guidelines and that I use question-types to formulate specific guidelines. So we have a *general, unified* strategy for avoiding non-significant questions: guidelines with certain conditions in the antecedent and questions of certain types as output. Khalifa rightly stresses that we have to avoid nonsensical why-questions such as "Why is George W. Bush president rather than George W. Bush president?' in which topic and foil are identical (2004, p.46). The guidelines guarantee this. Khalifa's semantic and methodological strategies are meant to avoid too much variation. The different question-types distinguished above avoid excessive variation between topic and foil in two ways:

(1) the objects are identical or similar (as they were in the examples of the crop fields and the American and British armies)
(2) (2) P and P* are mutually exclusive, which entails that they belong to the same family of predicates.

In this way, the question types make the semantic and methodological strategy superfluous, as they enable us to avoid the green cheese problem without them. The underlying idea of the semantic strategy (shared reference) is incorporated in (1), while the fact that P and P* belong to the same family of predicates implies that they can be

measured by the same methods (so the methodological strategy is also incorporated in an indirect way).

The second additional difference is that, while Khalifa only considers expectations in his epistemic strategy, the approach advocated here also includes two guidelines based on ideals. As we saw in the examples of I-type and I'-type questions, sometimes the relevant contrast is between a state of affairs which we know to be true and one which we *desire* to be true; i.e. between a factual and an ideal situation, rather than an unexpected one. Thus, my proposal is more complete because it adds this important element (Khalifa could have included this idea in his epistemic strategy, but he did not).

Finally, with respect to the practical red herring task (helping scientists to avoid uninteresting answers) our solution is more precise than the one proposed by Khalifa. The strategies he offers for this task here are quite general:

> Generally, the relevance of answer to a why-question is determined by such considerations as its appeal to a causal mechanism, its ability to diffuse competing explanations, and the ease with which its variables can be isolated. (2004, p. 52)

The guidelines are more specific because they specify exactly what kind of causal information a significant answer should contain. This is possible because of the use of question-types and epistemic interests.

Chapter 13
Applying the pragmatic approach to how-questions

13.1 Introduction

Most of the questions we have been considering so far are why-questions. However, as I already pointed out in section 5.3, how-questions are legitimate explanation-seeking questions too, as we have seen over and over again when discussing model-explanations (whether of a functional or mechanistic kind). Moreover, even outside the narrow area of model explanation, interesting how-questions can arise.

To see this, let us briefly consider a classic in the literature of psycholinguistics: Chomsky's explanation of the capacity to acquire language, as demonstrated by children. Up to the late fifties, behaviourism was still dominant in psychology. Skinner's explanation of language avoided all references to mental processes: language acquisition was understood as the acquisition of a certain set of behavioural dispositions, which (under normal circumstances) included the tendency to utter phrases like "poke up the fire" when a room is cold, or "slow down!" when being driven at high speed along a stretch of treacherous cliffs. But how do we acquire these dispositions? Two things are important: the features of our current surroundings as they present themselves, and the history of reinforced behaviour (e.g. the parents rewarding the child when it uses expressions correctly). So the Skinnerian answer to the basic how-question: "

Q1 How do children acquire the use of language?"[92]

was couched in terms of children receiving specific environmental input and connecting it with a particular type of output through operant conditioning.

However, as is widely known, Chomsky launched a devastating attack on this behavioural explanation (1959). Specifically, he argued against the idea that operant conditioning could ever give rise to sets of dispositions that are sufficient to account for a person's verbal behaviour. Even acquiring the sets of disposition needed for the use of a simple word like 'cheese' would require a lengthy and severe training on the behaviourist account. A child would need to hear the word in affirmative and negating sentences, in questions and answers, as subject and object, in plural and singular, modified by adjectives etc. This, however, is a highly implausible account of how children acquire language. Children seem able to master, in a relatively short span of time, the usages of a word such that they are able to produce a sheer inexhaustible range of possible sentences. This range cannot be accounted for in terms of mere operant conditioning histories. This famous poverty-of-the-stimulus argument was reason to reject Skinner's answer to the question phrased above.

However, although it ruled out Skinner's answer to the how-question posed above as incompatible with empirical data, the poverty-of-the-stimulus argument did not undermine the scientific validity of the question itself. In fact, it led Chomsky to pose what one might call an *exclusive* how-question:

Q2) How is language acquisition in children possible, given the poor verbal input?

This type of question contrasts two state of affairs that are both known to be true but which seem to exclude each other. The direct reason to pose this question is an epistemic interest we already encountered in the previous chapter: the intellectual surprise brought on by the apparent exclusion. However, a second wish provoked by the apparent exclusion is the desire to end this situation: the epistemic interest of *consistency* is involved. The cognitive dissonance that stems from entertaining contrary opinions should be resolved, and so the answer should be such that it shows why the exclusion is only an apparent exclusion.

Chomsky tried to account for language acquisition in light of the poverty of the stimulus by positing an innate, universal grammar (generative or transformational

[92] Here we have an example of a legitimate non-contrastive explanation-seeking question; hence, I reject van Fraassen's second claim as listed at the end of section 3.2.

grammar) involving a finite set of recursive rules which are applied to phrase structures on a phrase structure tree, such that an infinite number of possible sentences can be formed. Chomsky's answer could be rendered as follows:

Language acquisition in children is possible even though

1. They receive only poor stimulus, because
2. They possess an innate set of recursive rules that allows them to manipulate the syntax of sentences beyond cases they have already experienced.

This answer shows why the exclusion is only apparent, because it reveals the hidden assumption: language acquisition and poor stimulus only exclude each other if we think of the former as being the result of operant conditioning. Thus, the answer serves the epistemic interest of *consistency*.

13.2 Accounting for how-questions[93]

Thus, scientists ask how- as well as why-questions. Given this situation, it might prove illuminating to explore the relation between these two. To do this, we must first ask ourselves whether how-questions are importantly different from why-questions. It may be that how-questions are reducible to why-questions, to the effect that how-questions can be restated as why-questions without loss of meaning, as Salmon once claimed (1984 p. 10).[94]

However, there are strong intuitions against this view. Questions like 'Why do bears hibernate' and 'How do bears hibernate?' simply do not mean the same thing. One can easily see this if one takes a moment to consider some possible contrast classes these questions might presuppose. In the former question, the contrastive clause might be: '...rather than maintain their normal sleeping habits throughout winter'; which would never do as a contrast class for the second question. And as the contrast classes can be different for why- and how-questions, so can the class of appropriate answers that could be given to such questions.[95]

[93] With some slight alterations, the following three sections have been published in Gervais 2012b

[94] It appears though that he later changes his mind about this (Salmon 1989 p. 137-138).

[95] See Faye 1999, 2007 for a detailed version of this argument.

If we assume then that how-questions are not only used by scientists, but also genuinely different from why-questions, a second question arises: Why bother with how-questions at all? If how-questions are not reducible to why-questions, then what reasons do we have to study them (assuming that why-questions are themselves unproblematic)? Perhaps this irreducibility is a sign that we should simply ignore them. Of course, from our previous examples, we know that if we want to understand the explanatory practices of cognitive scientists, we cannot afford to take this view: it is simply a matter of fact that scientists ask how-questions. However, it is conceivable that these scientists are simply wrong. In what follows, I will give two arguments against this view, i.e. in support the value of how-questions. In section 13.3, I will show that the knowledge yielded by answers to how- and why-question can be *complementary*, and that how-information can enjoy epistemological priority over why-information. In section 13.4, I will show that the answers to how-questions can lead to interesting new why questions.

13.3 The importance of how-questions #1

In some cases, the answers to how- and why-questions can *complement* each other. In medicine for example, knowledge about how bodily functions such as metabolism or growth are normally performed, may indicate where to look when such functions are impaired in a group or individual. Diagnostic reasoning in engineering contexts, where knowledge of the normal functioning of artefacts is used to analyse cases of malfunction (e.g. Bell, Snooke & Price 2007), provides another example.[96] Only if we understand how the process is normally carried out, can we understand the reason why it doesn't apply for some individual or group of individuals. In such cases, how-knowledge is complementary to why-knowledge.

Again, in the cognitive sciences, this situation is common. It has to do with the fact that an important type of explanandum in these disciplines are cognitive capacities. Now it is possible of course to ask a why-question about capacities. Generally speaking, if we take a system *S* and a capacity *C*, one could ask:

Q3) Why does *S* perform *C*?

[96] Thanks to Dingmar van Eck for suggesting this example.

Or, in generalized form:

Q4) Why do $S^{1...n}$ perform C

Presumably, this is a legitimate explanation-seeking question. To illustrate Q4, recall the example of the capacity to see depth. It is possible and indeed scientifically interesting to ask why humans can see depth, and this question might be answered in terms of evolutionary benefits of the capacity in question. However, in some cases, for example when a person has lost the capacity to see depth because of some injury, information about the evolutionary history of that capacity in humans is not what we need. In fact, elaborating upon this example, it seems we need to know why *this individual*, has lost the capacity to see depth. And so we might ask, for therapeutic purposes:

Q5) Why does S not perform *C*?

That is, we might ask: "Why does Jones not see depth?", which conforms to the schema of Q5. But in order to know the answer, we need to know *how* the capacity to see depth is *normally* realized in humans. If we understand how the capacity is realized other things being equal, this knowledge might suggest ways to repair the impaired capacity. Thus we arrive at the question:

Q5) How do S1...n perform C?

In short, in the therapeutic or preventive contexts, *why-questions about individuals can sometimes only be answered if we have first answered how-questions about groups*. The how-story about the group can give you the why-story about an individual: these stories complement each other. Hence, how-questions can be epistemologically prior to why-questions, as the answer to the former can guide the search for an answer to the latter. This is one way in which how-questions can be valuable.

13.4 The importance of how-questions #2[97]

This last point (that a how-story about the group can give you the why-story about an individual) can be fleshed out in terms of the pragmatic approach by applying it to model explanations, where model explanations are seen as answers to how-questions. Doing this will allow us to see a second sense in which how-questions are valuable: they can lead to interesting why-questions. This means applying the approach to how-questions. Let us start with a how question about an individual system performing a capacity:

Q6) *Plain fact*: How does S perform capacity C?

To stick to our example, the question: 'How does Jones realize the capacity to see depth?' conforms to Q6. I call this a plain fact question, because it is non-contrastive. The answer to this question might be:

Plain fact answer: By means of mechanism M

Presumably, the answer to the question why Jones sees depth will refer to incoming light, his visual pathway (eyes, optic nerves, lateral geniculate nucleus, visual cortex etc.), monocular clues (occlusion, familiarity, motion parallax etc.), and binocular clues (retinal disparity). Thus, M stands for a description of a complex, highly inter-level mechanism.

It is important to note here that just because we are now including how-questions, this does not mean we rule out why-questions. Thus, the answer just given might prompt the following why question:

Q7) Why does system S perform C by means of M?

The answer to this question could refer to the system's evolutionary past, or it could be inferred or assumed from its similarities to other systems. That is, if we want to know why Jones perceives depth in the manner described above, we are likely asking for an evolutionary story about the beneficial qualities of that manner (as opposed to other

[97] The following section is taken from Gervais & Kosolosky (2012).

mechanisms by which the same capacity could be realized). Now consider the generalized version of the plain fact question-type:

Q8) How do systems $S^{1...n}$ perform capacity C?

and its answer:

Through an M-like mechanism.

Q6 asks how a specific system performs a certain capacity, and it is answered by providing a model of the responsible mechanism. If we have answered this question-type for enough systems like S, then, based on induction, we have a generalized answer to Q8: all systems like S perform capacity C through M-like mechanisms. In fact, and this is where we start interpreting our answers, the answer to Q8 can then be taken to read something like this: S-like systems *normally* realize C through M-like mechanisms: humans normally perform the capacity to see depth by means of the monocular clues and binocular clues just mentioned. Of course, the more diverse the range of systems we want to generalize across, the more abstract our model of M has to be.

Now if we know the answer to Q8, this gives us a something like a covering-law answer to Q7: a particular system S realizes capacity C through mechanism M, because all systems like S realize C through M. Similarly, the generalized version of Q7:

Q9) Why do systems $S^{1...n}$ perform capacity C through an M-like mechanism?

could be answered in the CL fashion if we know systems $S^{1...n}$ to be a subset of some still larger set of systems that all realize C through M. However, the model covering this still larger subset will be even more abstract, perhaps to the extent of becoming trivial.

Until now, we have just considered plain fact questions and their answers, but of course, contrastive questions like the ones we encountered in the previous chapter may play a role. We can easily see this if we consider a system S that behaves out of the ordinary, in that it fails to perform a capacity other S-like systems do perform (for example if Jones has lost the capacity to see depth). Here, we have an unexpected contrast on our hands, between the behaviour of an individual system (Jones) and the behaviour of other systems belonging to the set (the human species) of which that system is a member. Here, we can ask an I'-type question. As you will recall (see section 12.4.1), the general format of an I-type question is: "Why does *x* have property P, while *y* has the ideal property P*?" Tweaking a bit, we get:

Q10 Why does S fail to perform C, while $S^{1...n}$ do perform C?"

Or, in terms of the example:

Q11 Why does Jones fail to see depth, while other humans do not?

Of course, in everyday and medical contexts, the contrastive part of this question usually left implicit.

Sometimes, we may find that although a system performs the capacity we would expect it to perform given its membership of a set of systems with the ability to perform that capacity, it does so in a different way. A particularly interesting example here is the brain's ability to reroute its operations after injury, so as to restore a previously impaired capacity. We might know that a particular lobe, normally involved in performing a capacity C, is damaged in an individual, impairing C. However, after a while we might observe that C is somehow restored. Here, we might ask:

Q12 How does S perform C, given that M is not available?

Let us call this an R-type question. Here, we contrast the mechanism realizing C in an individual with the mechanism responsible for realizing C in other S-like systems, because we know there must be a difference. Knowing what this difference consists of can help us to understand other cases in which a previously impaired capacity is restored: some rerouting or reorganizing solutions may be more common than others. In turn, this may prompt still other why-questions, as we would like to know why some strategies for restoring capacities are more common than others. Thus, it appears that the answers to how-questions can lead us to formulate new why-questions.

Part 5

Conclusion

Chapter 14 Conclusion and future prospects

14.1 Results for robust pluralism.

It is time now to take stock and summarize the main results of this dissertation. As I have stressed in section 4.1, we should distinguish the pragmatic approach to explanation, which is a meta-philosophical theory on how to study scientific explanation, from robust pluralism (the domain-specific descriptive and normative claims), which is the philosophical position resulting from applying the pragmatic approach. Let me start this concluding chapter by looking at the progress we made with robust pluralism; in section 14.2, I will do the same for the pragmatic approach. Finally, in section 14.3 I will conclude this dissertation by sketching some outlines for future research.

Recall that in section 3.4, we defined robust pluralism as the combination of the following four claims:

A) There are no general exclusion rules with respect to scientific explanations.

B) There are no general preference rules with respect to scientific explanations.

E) There are local exclusion rules with respect to scientific explanations.

F) There are local preference rules with respect to scientific explanations.

Where A and B capture minimal explanatory pluralism, and E and F add the robustness we are looking for. So how robust is robust pluralism?

In my view, this depends on two factors. The first is the amount of local exclusion and preference rules – as we add more, robustness increases. Here is a list of the *local*

exclusion and preference rules I am confident enough to propose, based on the material and the examples covered in this dissertation:

1- (*Exclusion rule*) In dynamical cognitive science, DN explanations do not exist (they are non-deductive causal CL explanations using a default rule).

2- (*Preference rule*) In biology, if one wants to understand contrasts between the capacities of different species while the mechanism is largely unknown, DN explanations are superior to mechanistic explanations.

3- (*Exclusion rule*) In biology, if there is only evidence from hypothetico-deductive experiments available (and consequently bottom-up experiments are impossible), mechanistic explanations cannot be given.

4- (*Preference rule*) In cognitive science, both functional and mechanistic models are superior to merely phenomenal models.

5- (*Preference rule*) In artificial intelligence, if performance is the main interest, and the model in question is outperformed by a relevant set of natural systems, plausibility is a virtue.

6- (*Preference rule*) In artificial intelligence, if performance is the main interest, and the model in question outperforms a relevant set of natural systems, plausibility is a vice.

These rules are argued for in part II and III of this dissertation, and I will readily admit that to anyone who has not read these parts, some of these rules will be hard to understand, or indeed seem implausible. However, since I have followed the bottom-up methodology of the pragmatic approach as described in section 1.4, all these rules are backed up by examples from the scientific literature.

In any case, adding more of these local rules boosts the robustness of our pluralism. However, not all rules boost it to the same degree. Here I come to the second factor: how interesting are the local exclusion rules? In turn, the answer to this question will also depend on (at least) two factors. One, there is what one might call the controversy factor. The idea here is that if a local exclusion or preference rule goes against the current fashion among theorists in the field, then this makes it interesting for both scientists and philosophers. From this point of view, rule 2 is very interesting, for as I have shown in section 8.1, the presence of DN explanations in biology is highly contested. Rule 4 will strike most readers as not so interesting (although it would be controversial for anyone who thinks of functional models as phenomenal models).

The second factor determining how interesting a local exclusion or preference rule is has to do with its scope: *how* local is the local rule? As they are stated above, the rules

are claimed to apply to specific scientific disciplines. No doubt some of them (especially the preference rules) will, upon investigation, turn out to have a wider scope, but this is of course something that can only be argued for by illustrating them with examples of explanations taken from the scientific literature itself.

14.2 The results for the pragmatic approach

14.2.1 Epistemic interests

As stated in the first chapter, the pragmatic approach is a means of studying scientific explanation by taking into account the epistemic interests behind the explanation, and the format of explanation-seeking questions and their answers. In chapter 12, we have seen that the epistemic interests can be used to formulate heuristic guidelines. This is important, because this heuristic, forward-looking use of the interests represents a genuine enlargement of the pragmatic approach, which previously used interests mainly to analyse existing explanations, resulting in local preference and exclusion rules rather than guidelines. With this in mind, it is useful to take a moment to make an inventory of these interests.

The epistemic interests we have come across so far can be ordered into two categories, theoretical and practical. Here are the two lists, together with a brief explication (in some cases, the meaning is rather obvious):

Theoretical epistemic interests

Understanding	Basic desire to know; sheer intellectual curiosity, etc.
Fruitfulness	Desire to obtain an explanation that is conducive to furthering future research (by suggesting new experiments or new methodological approaches, by leading to new explanation-seeking questions, etc.).[98]

[98] Thus, my use of the notion 'fruitfulness' as a particular epistemic interest should not be confused with Carnap's use of the same notion as a property of scientific concepts (see note 28). One might say though that Carnap's conception of fruitfulness is a special instance of my more general concept: being useful in

Surprise	Curiosity brought on by a perceived difference between what we know to be the case and what we expected to be the case.
Unification/consistency	The interest of making the (occurrence of) the explanandum 'fit' with our other opinions or established theories (web of belief). Often, unification is a continuation of surprise.

Practical epistemic interests

Control/manipulation	Basic desire to be able to control, manipulate or intervene upon, some process, e.g. in experimental setups.
Prediction	The aim of obtaining statements (e.g. from derivation of other statements) describing events that follow given some initial state.
Prevention	Desire to stop a certain state of affairs from occurring in the future.
Improvement	Desire to bring about a state of affairs which we deem preferable to the actual state of affairs.
Duplication	Interest to make a model duplicate the performance of a certain capacity exhibited by another (natural or artificial) system.
Performance	Aim to make a model perform a capacity better (faster, more stably, with greater success rate etc.).
Responsibility	Desire to obtain an explanation attributing moral and/or legal responsibility.

There are three observations to be made about this list (some of the following points we have come across already). First, they can relate to each other in various ways. They can act complimentary, or they can pull in opposite directions; one can act as a subsidiary goal to another (as we have seen, in some of the models used in artificial intelligence, duplication acts as an intermediate aim on the route to maximum performance), and some of the interests are specific instances of others (prevention, for example, is a specific variant of the general interest of control). Second, the distinction between theoretical and practical epistemic interests is not absolute. For example, in some contexts, such as when considering Popper's falsifiability demarcation criterion (Popper

formulating new universal statements is *one* way in which a concept, and therefore an explanation citing that concept, can be conducive to further research.

1963), prediction might well be interpreted as a theoretical interest. Finally (and this should come as no surprise), the two lists are not exhaustive.

14.2.2 Question-types

Realizing what interests are present in a given case of scientific explanation is important, for as we have seen it has an impact on what kind of explanation-seeking question is appropriate. Although some of these explanation-seeking questions have already been formulated in earlier publications by researchers working within the pragmatic approach, it may be helpful to list the ones we have considered so far in this dissertation, together with some of the epistemic interests they can serve:

Plain fact I	Why X?	Understanding
Plain fact II	How does system S perform capacity C?	Understanding, improvement
Plain fact III	How do systems $S^{1...n}$ perform capacity C?	Understanding, improvement
I-type	Why does *x* have property P, rather than the ideal property P*?	Improvement, prevention, attributing responsibility
I'-type	Why does *x* have property P, while *y* has the ideal property P*?	Improvement, prevention, attributing responsibility
E-type	Why does *x* have property P, while objects $y_1,..., y_n$ have property P*?	Surprise, consistency
T-type	Why does *x* have property P* at *t'*, while it had property P at *t*?	Surprise, consistency
R-type	How does system S perform C, given that mechanism M is not available?	Surprise, improvement

Again, some observations can be made here. First, the questions all ask for explanations in their own right, but they can be asked at different phases of the more broad explanatory process a scientist is engaged in (for example, the answer to the R-type question presupposes an answer to the plain fact III question). Second, the epistemic interests listed in the third column are only chosen because they are the ones which happened to be operative in the examples I used to illustrate the given question-types with. It is very possible that other interest may be at work in a given case (for example, a T-type question may be asked to determine moral responsibility; the plain fact II

question might be asked to add richness to a model etc.). Finally, as with the epistemic interests, this list is not exhaustive.

14.3 Future prospects

So how could the pragmatic approach be further developed? How can the philosophical position of robust pluralism be further substantiated? I will conclude this dissertation by offering two general guidelines and one more concrete proposal.

Regarding the pragmatic approach, one could add to the stock of epistemic interests and question-types listed above. As we have seen in chapter 12, these question-types can be used to formulate heuristic guidelines conforming to the general schema: 'If conditions c_1, ..., c_n, hold then question X is a significant explanation-seeking question.' We have already formulated such guidelines for I-, I'-, E- and T-type questions, but it is of course possible to formulate new guidelines based on the other question-types. Moreover, as how-questions come within our grasp, the potential field of application is substantially enlarged. How-questions and model explanations play an important role, not only in cognitive science and biology, but also in the life-sciences at large. Here, the recipe seems to be to analyse ever more case studies of scientific explanation, from ever more disparate sources, extending the toolbox as one goes along. It is a matter of staying true to the idea that one can express explanations as answers to certain questions, and that these questions are always asked with certain aims in mind. Thus, although one expands the toolbox, the meta-philosophical position on how to study explanations, in broad lines remains the same.

The task of further substantiating the position of robust explanatory pluralism goes hand in glove with the further development of the pragmatic approach. Again, it is important here to realize that the list of section 14.1 is quite obviously incomplete, and necessarily so. Here, to add to the robustness of robust pluralism means to carefully scrutinize the results of the case studies to which you have applied the pragmatic approach, and then adding to the list of domain-specific or local exclusion and preference rules – preferably ones that are interesting in the sense of applying to multiple domains, or going against the philosophical fashion.

Finally, there is one concrete issue to which the pragmatic approach we have developed so far might be applied. I am thinking here of inter-level explanations, where levels are understood as being constituted by scientific vocabularies. This would be important for two reasons. First, inter-level explanations are important because they are numerous. For example, in the cognitive sciences, a psychological state like depression might be explained in terms of a neurophysiological condition like reduced

serotonin levels. Similarly, in the earth sciences, an explanation of the shape of the continents might include references to the chemical properties of magma. Another important motivation to apply the pragmatic approach to inter-level explanations has to do with the two disciplines that have featured prominently in this dissertation: cognitive science and biology. Both are comprised of many sub-disciplines located at many different levels of analysis, and both can at times be thoroughly interdisciplinary. In particular, philosophers who study cognitive science tend to be highly interested in inter-theoretic relations and cross-disciplinary explanations. The reason for this is at least in part historical: in the debate about cognitive science and inter-theoretic relations, one can discern the last vestiges of the venerable mind-body problem. In effect, the discussion about the relation between psychological and neurological theories constitutes the latest stage of a transition from metaphysics to philosophy of science that has been going on since at least the mid-twentieth century (Bickle 1998, 2003; Chemero & Silberstein 2008).

This last point is especially important, because the explanatory pluralism resulting from the pragmatic approach has already been argued for in the context of cognitive science and inter-theoretic relations (Bechtel & McCauley 1999; McCauley 1996, 2007; McCauley & Bechtel 2001). The four claims making up robust pluralism can be reformulated to apply specifically to inter-level explanations:

A*) There are no general exclusion rules with respect to grain size.
B*) There are no general preference rules with respect to grain size.
E*) There are local exclusion rules with respect to grain size.
F*) There are local preference rules with respect to grain size.

In terms of the debate in the philosophy of mind, these claims combined rule out a number of general positions: eliminativism is ruled out by A*, classic (Nagelian) reductionism and explanatory emergentism by B*. Of course, the history of science testifies of successful local reductions and eliminations, and these are not ruled out. The idea would then be that by analysing case studies of inter-level explanations one could show that grain size as such depends on contextual factors rather than metaphysical considerations. This is especially true in the case of mechanistic models, which typically proceed back and forth across many levels of description. Once the choice for the target or explanandum capacity has been made, there is no principled way to restrict ourselves to a particular level: we simply go wherever the causal and constitutive pathways take us.

Besides extending the explanatory pluralism to the issue of grain size however, there is another important point to take away from this brief consideration of inter-level explanations. If explanations are answers to why- or how-questions, then inter-level

explanations are answers to why-questions that are couched in a different scientific vocabulary: psychological questions receiving neurological answers, behavioural questions receiving biochemical answers, etc. That is, the relevance relation R, though certainly in need of further constraints as we have seen in chapter 3, cannot be taken to imply terminological uniformity between the explanation-seeking question and the answers, nor between competing answers. That is, the lesson here is that answers may be relevant to a question, despite being couched in a different scientific vocabulary, and that two answers may be in competition with each other, even though, again, both employ (radically) different vocabularies. This would mean taking issue with a view defended by Hardcastle:

> So then, if neuroscientific explanations are going to compete with psychological ones with respect to a set of questions, the contrast class for the two domains questions [sic] would have to correspond. But because psychology and neuroscience operate in such different academic environments, prima facie it is doubtful that the contrast classes do correspond. At least, we would need an argument that the contrast classes are importantly similar (Hardcastle 1998, p. 16-17).

I believe that the general argument Hardcastle asks for is not on the cards – the competition she refers to is a fact, whether we can devise such an argument or not. Rather, in the spirit of the pragmatic approach, one would need to show how certain epistemic interests can lead scientists to search for answers outside the domain from which the explanation-seeking question happens to originate.

Summary in Dutch

Verklaringen in de cognitiewetenschappen en de biologie. Mechanismen, Wetten en hun verklarende deugden.

Er zijn vermoedelijk weinig onderwerpen in de recente geschiedenis van de wetenschapsfilosofie die meer stof doen opwaaien dan wetenschappelijke verklaring. Het is inmiddels goed gebruik om een overzicht van de verschillende theorieën over verklaring die in de tweede helft van de twintigste eeuw zijn geformuleerd, aan te vangen met Hempels deductief-nomologische (DN) model, volgens welke het proces van verklaring bestaat uit het deduceren van het explanans (of liever: een propositie die het explanans beschrijft) uit ten minste één natuurwet en een aantal randvoorwaarden. Hoewel het model van Hempel definitief de status van een klassieker heeft verkregen, zijn er met het verstrijken van de tijd veel (soms schijnbaar onoverkomelijke) bezwaren tegen geformuleerd. In dit licht is het weinig verrassend dat het DN model al spoedig enkele rivalen naast zich moest dulden. Belangrijke alternatieve theorieën over verklaring zijn bijvoorbeeld Salmons causaal-mechanistische (CM) model en Kitchers unificationisme (U). Op hun beurt hebben ook deze alternatieven te kampen gekregen met tal van tegenwerpingen, zodat er van een onbetwiste winnaar in dit debat vooralsnog geen sprake lijkt te zijn.

Wie deze situatie overziet, kan zich moeilijk aan de indruk onttrekken dat wetenschappelijke verklaring een zodanig heterogene, dynamische aangelegenheid is, dat zij simpelweg elke poging om haar in termen van één model te vangen weerstaat. Zo ontstaat een pluralistisch beeld: verklaringen kunnen causaal of niet causaal zijn, ze kunnen refereren aan natuurwetten of aan statistische generalisaties, ze kunnen unificerend zijn of niet unificerend. Wellicht moeten de verklaringstypen die tot nu toe in de literatuur beschreven zijn, gezien worden als evenzovele stukken gereedschappen in een gereedschapskist, die, wanneer nodig, tevoorschijn gehaald kunnen worden om concrete wetenschappelijke verklaringen filosofisch te duiden, zonder daarmee aanspraak te kunnen maken op algemene geldigheid.

Dit verklaringspluralisme lijkt op het eerste gezicht een redelijke positie. Toch is het bij nader inzien filosofisch gezien weinig aantrekkelijk. Als er geen enkele algemene bewering over verklaring gedaan kan worden zonder op tegenvoorbeelden te stuiten, dan lijkt de filosoof zich tevreden te moeten stellen met het analyseren van een oneindige reeks concrete voorbeelden van verklaring. Volgens deze vorm van pluralisme is alles geoorloofd – het is een over-tolerante vorm van pluralisme.

De vraag die nu voorligt is de volgende: Is het mogelijk om enerzijds te erkennen dat er veel verschillende vormen van verklaring naast elkaar bestaan, maar anderzijds toch iets informatiefs over verklaring in het algemeen te zeggen? Dat wil zeggen: Is het mogelijk om tot een robuust soort van pluralisme te komen?

Dit laatste is precies de doelstelling van de *pragmatische benadering* van verklaringen, een meta-filosofische manier om verklaringen te bestuderen die ontwikkeld werd in het Centrum voor Logica en Wetenschapsfilosofie aan de Universiteit van Gent. De pragmatische benadering bestudeert verklaringen door ze te interpreteren als antwoorden op waarom-vragen in de vorm 'Waarom X (en niet Y)?'. Het idee is nu dat deze vragen gesteld worden met het oog op het verwezenlijken van specifieke doelstellingen of epistemische interesses. Deze epistemische interesses zijn niet alleen van invloed op de vorm van de waarom-vragen, maar dienen ook als toetssteen om de relevantie en het succes van de *antwoorden* op die vragen te evalueren – dit alles geïllustreerd aan de hand van concrete voorbeelden van verklaringen uit de wetenschappelijke literatuur. Kort gezegd is het doel van de pragmatische benadering om op deze manier tot domein-specifieke (in plaats van algemene) descriptieve en normatieve beweringen over verklaringen te komen.

Dit proefschrift bevat de bijdragen die ik in de loop van mijn doctoraat aan dit project geleverd heb. Specifiek heb ik geprobeerd om drie lacunes in de pragmatische benadering te verhelpen. Twee van deze lacunes hangen samen met het feit dat de pragmatische benadering tot nu toe niet of nauwelijks is toegepast op verklaringen met behulp van *modellen*. Dit kan als een tekort worden aangemerkt, al is het maar omdat modellen figureren in de verklaringspraktijk van talrijke wetenschappelijke disciplines, met name in de zogenaamde levenswetenschappen. Dit algemene tekort heb ik getracht te ondervangen door twee vragen te stellen: Hoe verhouden model-verklaringen zich tot andere vormen van verklaringen? En wat zijn eigenlijk de verschillende eigenschappen die modellen kunnen hebben – dat wil zeggen, welke epistemische interesses kunnen er eigenlijk ten grondslag liggen aan modellen? Deze twee vragen heb ik geprobeerd te beantwoorden aan de hand van voorbeelden van verklaringen uit de cognitiewetenschappen en de biologie. Tot slot is er nog een derde punt waarop de pragmatische benadering reflectie behoeft: hoewel de epistemische interesses en de rol die zij spelen in de totstandkoming van antwoorden op waarom-vragen veel aandacht hebben gekregen, zijn de vragen zelf relatief onderbelicht gebleven. Vragen die hier een rol spelen zijn: Wat voor verschillende soorten vragen kunnen er bijdragen aan de

bevrediging van welke epistemische interesses, hoe hangen die typen vragen samen, en zijn naast waarom-vragen ook hoe-vragen niet belangrijk voor het wetenschappelijke bedrijf? Dit laatste punt is extra relevant wanneer we ons bezig houden met modellen, die in de biologie en cognitiewetenschap immers vaak de vorm aannemen van beschrijvingen hoe een (natuurlijk of kunstmatig) systeem een bepaalde biologische of cognitieve capaciteit realiseert.

Deze drie onderwerpen komen aan bod in respectievelijk deel II, III en IV van dit proefschrift – ze worden voorafgegaan door een deel waarin ik kort de belangrijkste theorieën over verklaring samenvat, de problematiek van het pluralisme zoals hierboven geschetst uiteenzet en de belangrijkste tendensen van de pragmatische benadering weergeef. Deel II is onder andere gewijd aan wat men een voorzichtige rehabilitatie van Hempels model in de cognitiewetenschap en de biologie zou kunnen noemen. Hier wordt betoogd dat hoewel de verklaringen in deze disciplines zich niet strikt conformeren aan Hempels strenge voorwaarden, ze desalniettemin enkele formele eigenschappen van DN verklaringen overnemen. Dit gaat in tegen de huidige consensus onder wetenschapsfilosofen, die Hempels model al lijken te hebben afgeschreven ten faveure van zogenaamde mechanistische verklaringen (een specifiek soort van model-verklaringen).

In deel III concentreer ik mij op drie eigenschappen die modellen kunnen hebben: plausibiliteit, gedetailleerdheid en prestatie. Een aantal mogelijke relaties die tussen deze eigenschappen kunnen optreden worden geanalyseerd door twee typen modellen met elkaar te vergelijken: functionele en mechanistische modellen. Hier beargumenteer ik onder andere dat plausibiliteit en gedetailleerdheid onafhankelijk van elkaar kunnen optreden en dat plausibiliteit in bepaalde contexten ondergeschikt kan zijn aan prestatie. Met name het eerste gaat in tegen de gedachte, nu gangbaar onder de belangrijkste aanhangers van mechanistische verklaringen, dat de verklarende kracht van modellen wordt bepaald door de mate waarin ze als beschrijving van het betreffende mechanisme gedetailleerd zijn.

In deel IV ten slotte presenteer ik de belangrijkste algemene bijdragen die ik heb geleverd aan de pragmatische benadering van verklaringen. Ik pas hier onder andere de pragmatische benadering toe op hoe-vragen en ga in op de verhouding tussen waarom- en hoe-vragen. Ook presenteer ik hier enkele concrete heuristische richtlijnen om irrelevante vragen en antwoorden uit te sluiten, om zo het soort over-tolerante pluralisme dat ik aan het begin van deze samenvatting schetste in te ruilen voor een meer verfijnde vorm dat ik robuust pluralisme noem. Ik eindig met een samenvatting van de belangrijkste resultaten (een opsomming van de domein-specifieke descriptieve en normatieve beweringen over verklaring die we op basis van het behandelde materiaal kunnen maken) en een korte slotbeschouwing over de toekomstperspectieven van de pragmatische benadering.

References

Adini, Y., Moses, Y. & Ullman, S. (1997). Face recognition: The problem of compensating for changes in illumination direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19: 721-732.

Baddeley, A. D. (1998). *Human Memory: Theory and Practice*. Boston: Allyn and Bacon.

Bailer-Jones D. M. (2003). When scientific models represent. *International Studies in the Philosophy of Science* 17: 59-74.

Baker, J. M. (2005). Adaptive speciation: The role of natural selection in mechanisms of geographic and non-geographic speciation. *Philosophy of Science* 36: 303-326.

Beatty, J. (1995). The evolutionary contingency thesis. In: G. Wolters & J. Lennox (eds.) *Concepts, Theories, and Rationality in the Biological Sciences*. Pittsburgh: University of Pittsburgh Press, pp. 45-81.

Beatty, J. (1997). Why do biologists argue like they do? *Philosophy of Science* 64 (proceedings): S432-S443.

Bechtel, W. (2006). *Discovering Cell Mechanisms: The Creation of Modern Cell Biology*. Cambridge: Cambridge University Press.

Bechtel, W. (2007). Reducing psychology while maintaining its autonomy via mechanistic explanations. In: M. Schouten & H. Looren de Jong (eds.) *The Matter of the Mind: Philosophical Essays on Psychology, Neuroscience and Reduction*. Basil Blackwell: Oxford, pp. 172-198.

Bechtel, W. (2008). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. New York: Routledge.

Bechtel, W. (2009). Looking down, around and up: Mechanistic explanation in psychology. *Philosophical Psychology* 22: 543-564.

Bechtel, W. & Abrahamsen, A. (2005). Explanation: a mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences* 36, 421-443.

Bechtel, W. & McCauley, R. N. (1999). Heuristic Identity Theory (or back to the future): the Mind-Body problem against the background of research strategies in cognitive neuroscience. In: M. Hahn & S. C. Stoness (eds.) *Proceedings of the 21th Annual Meeting of the Cognitive Science Society*. Mahway, NJ: Lawrence Erlbaum Associates, pp. 67-72.

Bechtel, W. & Richardson, R. C. (1993). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton NJ: Princeton University Press.

Bell, J., Snooke, N. A. & Price, C. J. (2007). A language for functional interpretation of model based simulation. *Advanced Engineering Informatics* 21: 398-409.

Berelson, R. B., Lazarsfeld, P. F. & McPhee, W. (1954). *Voting: A Study of Opinion Formation in a Presidential Campaign*. Chicago: University of Chicago Press.

Bickle, J. (1998). *Psychoneural Reduction: The New Wave*. Cambridge MA: MIT Press.

Bickle, J. (2003). *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Deventer: Kluwer.

Bigelow, J. & Pargetter, R. (1987). Functions. *The Journal of Philosophy* 84, 181-196.

Boden, M. (2006). *Mind as Machine: A History of Cognitive Science (vol. 2)*. Oxford: Oxford University Press.
Brandon, R. N. (1997). Does biology have laws? The experimental evidence. *Philosophy of Science* 64 (proceedings): S444-S457.
Brédart, S., Valentine, T., Calder, A. & Gassi, L. (1995). An interactive model of face naming. *Quarterly Journal of Experimental Psychology* 81: 361-380.
Bruce, V. & Young, A. (1986). Understanding face recognition. *British Journal of Psychology* 77: 305-327.
Burton, A. M., Bruce, V. & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology* 81: 361-380.
Campbell, A., Converse, P. E., Miller, W. & Stokes, D. E. (1960). *The American Voter*. New York: Wiley.
Carnap, R. (1950). *Logical Foundation of Probability*. London: Routledge and Keegan Paul.
Carnap, R. (1966). The value of laws: explanation and prediction. In: R. Carnap (1966), *Philosophical Foundations of Physics*, pp. 12-16. New York: Basic Books.
Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford: Clarendon Press.
Chaney, C. K., Alvarez, M. R. & Nagler, J. (1998). Explaining the gender gap in U.S. presidential elections, 1980-1992). *Political Research Quarterly 51* (2), 229-311.
Chemero, A. (2009). *Radical Embodied Cognitive Science*. Cambridge MA: MIT Press.
Chemero, A. & Silberstein, M. (2008). After the philosophy of mind: replacing scholasticism with science. *Philosophy of Science* 75: 1-27.
Chomsky, N. (1959). Review of Skinner's *Verbal Behavior*. *Language* 35: 26-58.
Clark, A. (1997). *Being There: Putting Brain, Body and World Together Again*. Cambridge MA: Bradford/MIT Press.
Clark, A. (1998). Time and mind. *The Journal of Philosophy* 95: 354-376.
Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. New York: Oxford University Press.
Clark, A. & Chalmers, D. J. (1998). The extended mind. *Analysis* 58: 10-23.
Contessa, G. (2007). Scientific representation, interpretation and surrogative reasoning. *Philosophy of Science* 74 (1): 48-68.
Contessa, G. (2010). Scientific models and fictional objects. *Synthese* 172: 215-229.
Costain, A. N. & Berggren, H. (1998). The gendered electorate. Paper presented at the annual meeting of the American Political Science Association, Boston, September.
Craver, C. F. (2001). Role functions, mechanisms, and hierarchy. *Philosophy of Science* 68: 31-55.
Craver, C. F. (2006). When mechanistic models explain. *Synthese* 153: 355-376.
Craver, C. F. (2007). *Explaining the Brain*. Oxford: Clarendon Press.
Cummins, R. (1975). Functional analysis. *The Journal of Philosophy* 72, 741-765.
Cummins, R. (1980). Functional explanation. In: N. Block (ed.), *Readings in the Philosophy of Psychology* vol. *1* (pp. 185-190). Cambridge MA: Harvard University Press.
Cummins, R. (2000). "How does it work?" versus "What are the laws?": Two conceptions of psychological explanation. In: F. C. Keil & R. A. Wilson (eds.), *Explanation and Cognition*. Cambridge MA: MIT Press, pp. 117-144.
D'Andrade, R. (1986). Three scientific world views and the covering law model. In: D. W. Fiske & R. A. Shweder (eds.) *Metatheory in Social Science*. Chicago: University of Chicago Press, pp. 19-41.
Da Costa, N. & French, S. (2003). *Science and Partial Truth: A Unitary Approach to Models and Scientific Reasoning.* Oxford: Oxford University Press.
Darden, L. (2005). Relations among fields: Mendelian, cytological and molecular mechanisms. *Philosophy of Science* 36: 349-371.
Darden, L. (2006). *Reasoning in Biological Discoveries: Essays on Mechanisms, Interfield Relations, and Anomaly Resolution*. Cambridge: Cambridge University Press.

Darden, L. & Tabery, J. (2009, September 9). Molecular Biology. In: E. N. Zalta (ed.) The Stanford Encyclopedia of Philosophy. Retrieved june 10, 2011 from http://plato.stanford.edu/archives/fall2010/entries/molecular-biology.

Dennett, D. C. (1978). Artificial intelligence as philosophy and as psychology. In: D. C. Dennett, *Brainstorms. Philosophical Essays on Mind and Psychology*. VT: Bradford Books, pp. 109-126.

Dennett, D. C. (1987). *The Intentional Stance*. Cambridge MA: MIT Press.

de Regt, H. & Dieks, D. (2005). A contextual approach to scientific understanding. *Synthese* 144, 137-170.

De Vreese, L., Weber, E. & Van Bouwel (2010). Explanatory pluralism in the medical sciences: Theory and practice. *Theoretical Medicine and Bioethics* 31: 371-390.

De Winter, J. (2010). Explanations in software engineering: The pragmatic point of view. *Minds and Machines* 20: 277-289.

Diamond, A. (1985). Development of the ability to use recall to guide action, as indicated by infants' performance on A-not-B. *Child Development* 56: 868-883.

Dowe, P. (2000). *Physical Causation*. Cambridge: Cambridge University Press.

Dowe, P. (2003). Causes are physically connected to their effects: Why preventers and omissions are not causes. In: C. Hitchcock (ed.), *Contemporary Debates in Philosophy of Science. Oxford*: Blackwell, pp. 198-206.

Faye, J. (1999). Explanation explained. *Synthese* 120: 61-75.

Faye, J. (2007). The pragmatic-rhetorical theory of explanation. In: J. Persson & P. Ylikoski (eds.), *Rethinking Explanation*. Dordrecht: Springer Verlag, pp. 109-118.

Feynman, R. (1986). *Appendix F- Personal Observations of the Reliability of the Shuttle*. Retrieved 12-02-2013 from: http://history.nasa.gov/rogersrep/v2appf.htm. [reprinted in Feynman 1988].

Feynman, R. (1988). *What Do You Care What Other People Think?* New York & London: W. W. Norton.

Fleissner, G., Holtkamp-Rötzler, E., Hanzlik, M., Winklhofer, M., Fleissner, G., Petersen, N. & Wiltschko, W. (2003). Ultrastructural analysis of a putative magneto receptor in the beak of homing pigeons. *The Journal of Comparative Neurology* 458, 350-360.

Fletcher, J. (1986). Report to the President: Actions to Implement the Recommendations of the Presidential Commission on the Space Shuttle Challenger Accident. NASA. Retrieved 21-02-2012 from: http://history.nasa.gov/rogersrep/actions.pdf.

Fodor, J. A. (1974). Special sciences, or the disunity of science as a working hypothesis. *Synthese* 28: 77-115.

Fodor, J. A. (1997). Special sciences: still autonomous after all these years. *Philosophical Perspectives* 11: 149-163.

Foster, R. G. & Kreitzman, L. (2009). *Seasons of Life: The Biological Rhythms that Enable Living Things to Thrive and Survive*. New Haven CT: Yale University Press.

Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy* 71: 5-19.

Frigg, R. (2010). Models and fiction. *Synthese* 172: 251-268.

Gardner, H. (1985). *The Mind's New Science. A History of the Cognitive Revolution*. New York, Basic Books.

Garfinkel, A. (1981). *Forms of Explanation*. New Haven: Yale University Press.

Garner, W. W. (1933). Comparative responses of long-day and short-day plants to relative length of day and night. *Plant Physiology* 8, 347-356.

Garner, W. W. & Allard, H. A. (1920). Effect of the relative length of day and night and other factors the environment on growth and reproduction in plants. *Journal of Agricultural Research* 18, 553-606.

van Gelder, T. (1995). What might cognition be, if not computation? *Journal of Philosophy* 91: 345-381.

Gervais, R. (2012a). Op het snijvlak van cognitie, wetenschap en filosofie: intertheoretische relaties in de twintigste eeuw. *Algemeen Nederlands Tijdschrift voor de Wijsbegeerte* 104: 21-38.
Gervais, R (2012b). Pragmatic approaches to explanation applied to the cognitive sciences. Two explanation-seeking questions compared. In: B. Van Kerkhove, T. Libert, G. VanPaemel & Marage, P. (eds.) *Logic, Philosophy and History of Science in Belgium II*. Koninklijke Vlaamse Academie van België, Brussel, pp. 131-138.
Gervais, R (2012c). Pragmatic aspects of inter-level explanations in the cognitive sciences. *Unpublished manuscript.*
Gervais, R. (In print). Explaining capacities: Assessing the explanatory power of models in the cognitive sciences. In: J. Meheus, E. Weber & D. Wouters (eds.). *Logic, Reasoning, and Rationality*. Dordrecht: Springer.
Gervais, R. & Kosolosky (2012). Understanding model explanations in biology: a pragmatic approach. Unpublished manuscript.
Gervais, R. & Looren de Jong, H. (In print). The status of functional explanation in psychology: Reduction and mechanistic explanation. *Theory & Psychology*.
Gervais, R. & Weber, E. (2011). The covering law model applied to dynamical cognitive science: A comment on Joel Walmsley. *Minds and Machines* 21: 33-39.
Gervais, R. & Weber, E. (2013a). Plausibility versus richness in mechanistic models. *Philosophical Psychology* 26: 139-152.
Gervais, R. & Weber, E. (2013b). DN Explanations in biology. *Manuscript submitted for publication.*
Gervais, R. & Wieland, J. W. (2012). Why don't effects explain? *Unpublished manuscript*.
Giere, R. (2004). How models are used to represent reality. *Philosophy of Science* 71 (supplement): S742-S752.
Giere, R. (2009). Why scientific models should not be regarded as works of fiction. In: M. Suárez (ed.), *Fictions in Science. Philosophical Essays on Modelling and Idealisation*. London: Routledge 248-258.
Giger, N. (2009). Towards a modern gender gap in Europe? A comparative analysis of voting behaviour in 12 countries. *The Social Science Journal 46*, 474-492.
Gijsbers, V. (2007). Why unification is neither necessary nor sufficient for explanation. *Philosophy of Science* 74: 481-500.
Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science* 69 (Supplement), S342-S353.
Glennan, S. (2005). Modelling mechanisms. *Studies in History and Philosophy Biological and Biomedical Sciences* 36: 443-464.
Godfrey-Smith, P. (2006). The strategy of model-based science. *Biology and Philosophy* 21: 725-740.
Greene, N. (n.d.). Space Shuttle Challenger Disaster – a NASA Tragedy. Part 2: The Space Challenger Aftermath, in: *About.Com (Part of the New York Times Company)*. Retrieved 21-02-2012 from: http://space.about.com/cs/challenger/a/challenger_2.htm.
Haken, HL, Kelso, J. A. S. & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics* 51: 347-356.
Hankinson, R. J. (1998). Cause and Explanation in Ancient Greek Thought. Oxford: Clarendon Press.
Hanley, J. R. (2011). An appreciation of Bruce and Young's (1986) serial stage model of face naming after 25 years. *British Journal of Psychology* 102: 915-930.
Hardcastle, V. G. (1998). On the matter of minds and mental causation. *Philosophy and Phenomenological Research* 58: 1-25.
Hausman, D. M. (1998). *Causal Asymmetries*. Cambridge: Cambridge University Press.
Haxby, J. V., Hoffman, E. A. & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences* 4: 223-233.
Hécaen, H. (1981). The neuropsychology of face recognition. In: G. Davies, H. Ellis and J. Shepherd (eds.) *Perceiving and Remembering Faces*. London: Academic Press, pp. 39-54.

Hempel, C. G. (1962). Deductive-nomological vs. statistical explanation. In: H. Feigl & G. Maxwell (eds.) *Minnesota Studies in the Philosophy of Science vol. 3*. Minneapolis: University of Minnesota Press, 98-169.
Hempel, C. G. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.
Hempel, C. G. (1966). *Philosophy of Natural Science*. Englewood Cliffs, NJ: Prentice Hall.
Hempel, C. G. & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science* 15: 135-175. Reprinted in Hempel 1965, pp. 245-290.
Hitchcock, C. (1995). Discussion: Salmon on explanatory relevance. *Philosophy of Science* 62: 304-320.
Inglehart, R. & Norris, P. (2000). The developmental theory of the gender gap: Women's and men's voting behaviour in global perspective. *International Political Science Review/Revue internationale de science politique 21* (4), 441-463.
Inglehart, R. & Norris, P. (2003). *Rising tide. Gender equality and cultural change.* Oxford: Cambridge University Press.
Kanwisher, N. G. & Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions – Royal Society. Biological Sciences* 361: 2109-2128.
Kaplan, D. M. (2012). How to demarcate the boundaries of cognition. *Biology and Philosophy* 27: 545-570.
Keeton, W. & Gould, J. (1986). *Biological Science*. New York and London: W. W. Norton & Co. (4th edition).
Kelso, J. A. S. (1995). *Dynamic Patterns: The Self-Organisation of Brain and Behavior*. Cambridge MA: MIT Press.
Khalifa, K. (2004). Erotetic contextualism, data-generating procedures and sociological explanations of social mobility. Philosophy of the Social Sciences 34 (1), 38-54.
Kitcher, P. (1981). Explanatory Unification. *Philosophy of Science* 48: 507-531.
Kitcher, P. (1985). Two approaches to explanation. *Journal of Philosophy* 82: 632-639.
Kitcher, P. (1989). *Explanatory unification and the causal structure of the world*. In: Kitcher & Salmon (1989).
Kitcher, P. (2001). *Science, Truth, and Democracy*. New York: Oxford University Press.
Kitcher, P. & Salmon, W. (1987). Van Fraassen on explanation. *Journal of Philosophy* 84 (6), 315-330.
Kitcher, P. & Salmon, W. (eds.) (1989). *Scientific Explanation*. Minneapolis: University of Minnesota Press.
Kyburg, H. (1965). Comment. *Philosophy of Science* 32: 147-151.
Leuridan, B. (2010). Can mechanisms really replace laws of nature? Philosophy of Science 3: 317-340.
Leuridan, B. (2012). Three problems for the mutual manipulability account of constitutive relevance in mechanisms. *British Journal for the Philosophy of Science* 63, 399-427.
Levelt, W. (1989). *Speaking: From Intention to Articulation*. Cambridge MA: MIT Press.
Lipset, S. M. (1960). *Political Man*. Baltimore: John Hopkins University Press.
Lipton, P. (1990). Contrastive explanation. In: D. Knowles (ed.) *Explanation and Its Limits*. Cambridge, Cambridge University Press, pp. 247-266.
Lipton, P. (1993). Making a difference. *Philosophica* 51: 39-54.
Machamer, P. K., Darden, L. & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science* 67: 1-25.
Manza, J. & Brooks, J. (1998). The gender gap in U.S. presidential elections: When? Why? Implications? *American Journal of Sociology 102* (5), 1235-1266.
Marr, D. & Nishihara, H. K. (1978). Visual information processing: Artificial intelligence and the sensorium of sight. *Technology Review* 81: 1-23.

McCauley, R. N. (1996). Explanatory pluralism and the coevolution of theories in science. In: R. N. McCauley (ed.) *The Churchlands and their critics*. Oxford: Blackwell, pp. 17-47.
McCauley, R. N. (2007). Enriching Philosophical Models of Cross-Scientific Relations: Incorporating Diachronic Theories. In: M. Schouten & H. Looren de Jong (eds.) *The Matter of the Mind: Philosophical Essays on Psychology, Neuroscience and Reduction*. Oxford: Blackwell Publishers, pp. 199-223.
McCauley, R. N. & Bechtel, W. (2001). Explanatory pluralism and the Heuristic Identity Theory. *Theory & Psychology* 11: 736-760.
McClelland, J. & Rumelhart, D. (1986). *Parallel distributed processing: Explorations in the micro-structure of cognition* (vol. 2). Cambridge, MA: MIT Press.
Mellor, D. H. (1988). On raising the chances of effects. In: J. Fetzer (ed.), *Probability and Causality: Essays in Honour of Wesley C. Salmon*. Dordrecht: Reidel, pp. 229-239.
Miller, G. A. (2003). The cognitive revolution: a historical perspective. *Trends in Cognitive Science* 7: 141-144.
Milner, B. (1966). Amnesia following operation on the temporal Lobes. In: C. W. M. Whitty & O. L. Zangwill (eds.) *Amnesia*. London: Butterworths, pp. 109-133.
Mitchell, S. D. (1997). Pragmatic laws. *Philosophy of Science* 64 (proceedings): S468-S479.
Mitchell, S. D. (2000). Dimensions of scientific laws. *Philosophy of Science* 67: 242-256.
Mitchell, S. D. (2003). *Biological Complexity and Integrative Pluralism*. Cambridge: Cambridge University Press.
Mueller, C. (ed.) (1988). *The Politics of the Gender Gap.* London: Sage.
Munkata, Y. (1998). Infant preserverative and implications for object permanence theories: A PDP model of the A-not-B task. *Developmental Science* 1: 161-184.
Murneek, A. E. (1948). History of research in photoperiodism. In: A. E. Murneek & R. O. Whyte (eds.), *Vernalization and Photoperiodism. A Symposium*. Waltham MA: Chronica Botanica Company, pp. 39-61.
Nagel, E. (1961). *The Structure of Science: Problems in the Logic of Scientific Explanation*. New York: Harcourt, Brace, and World.
Nozick, R. (1981). *Philosophical Explanations*. Cambridge MA: Harvard University Press.
Piaget, J. (1954). *The Construction of Reality in the Child*. New York: Basic Books.
Piaget, J. (1963). *The Origins of Intelligence in Children*. New York: W.W. Norton & company.
Piccinini, G. & Craver, C. (2011). Integrating psychology and neuroscience: functional analyses as mechanism sketches. *Synthese* 183: 283-311.
Popper, K. (1963). *Conjectures and Refutations*. London: Routledge and Kegan Paul.
Ramsey, J. (2008). Mechanisms and their explanatory challenges in organic chemistry. *Philosophy of Science* 75: 970-982.
Ramsey, W., Stich, S. P. & Garon, J. (1990). Connectionism, eliminativism, and the future of folk psychology. *Philosophical Perspectives* 4: 499-533.
Rhodes, G. (1985). Lateralized processes in face recognition. *British Journal of Psychology* 76: 249-271.
Risjord, M. (2000a). *Woodcutters and witchcraft: Rationality and interpretive change in the social sciences*. Albany, New York: State University of New York Press.
Risjord, M. (2000b). The politics of explanations and the origins of ethnography. *Perspectives on Science* 8: 29-52.
Rothstein, P., Henson, R. N., Treves, A., Driver, J. & Dolan, R. J. (2005). Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. *Nature Neuroscience* 8: 107-113.
Ruben, D. (1987). Explaining contrastive facts. *Analysis* 47: 35-37.
Ruben, D. (1990). Arguments, laws, and Explanation. In: D. Ruben, *Explaining Explanation*. New York: Routledge pp. 181-205.
Salmon, W. C. (1971a). Statistical Explanation. In: Salmon, W. (1971b), pp. 29-87.

Salmon, W. C. (ed.) (1971b). *Statistical Explanation and Statistical Relevance*. Pittsburgh: University of Pittsburgh Press.
Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton, NY: Princeton University Press.
Salmon, W. C. (1989). Four decades of scientific explanation. In: P. Kitcher & W. C. Salmon (eds.) *Scientific Explanation: Minnesota Studies in the Philosophy of Science (vol. XIII)*. MN: University of Minnesota Press, pp. 3-129.
Salmon, W. C. (1990). Causal propensities: Statistical causality vs. aleatory causality. *Topoi* 9 (2): 95-100.
Salmon, W. C. (1994). Causality without counterfactuals. *Philosophy of Science* 61: 297-312.
Salmon, W. C. (1998). *Causality and Explanation*. Oxford: Oxford University Press.
Schaffer, J. (2003). Causes need not be physically connected to their effects: the case for negative causation. In: C. Hitchcock (ed.), *Contemporary Debates in Philosophy of Science. Oxford*: Blackwell, pp. 207-226.
Scriven, M. (1962). Explanation, prediction, and laws. In: H. Feigl & G. Maxwell (eds.) *Scientific Explanation, Space, and Time: Minnesota Studies in the Philosophy of Science (vol. III)*. MN: University of Minnesota Press, pp. 170-230.
Shallice, T. & Warrington, E. K. (1980). Single and multiple component central dyslexic syndromes. In: M. Coltheart, K. E. Patterson & J. C. Marshall (eds.) *Deep Dyslexia*. London: Routledge and Kegan Paul, pp. 119-145.
Sober, E. (1997). Two outbreaks of lawlessness in recent philosophy of biology. *Philosophy of Science* 64 (proceedings): S458-S467.
Spirtes, P., Glymour, C. & Scheines, R. (1993). *Causation, Prediction and Search*. New York: Springer-Verlag. 2nd edition (2000) Cambridge: MIT Press.
Sterelny, K. (1990). *The Representational Theory of Mind: An Introduction*. Oxford: Blackwell.
Stouffer, S. (1955). *Communism, Conformity, and Civil Liberties.* New York: Doubleday.
Strevens, M. (2008). *Depth: An Account of Scientific Explanation*. Cambridge MA: Harvard University Press.
Sukthankar, G. (2000). Face recognition: a critical look at biologically-inspired approaches. Carnegie Mellon University, Pittsburgh PA. Technical report: CMURI-TR-00-04.
Suppes, P. (1970). *A Probabilistic Theory of Causality. Amsterdam*: North-Holland.
Tabery, J. G. (2004). Synthesizing activities and interactions in the concept of a mechanism. *Philosophy of Science* 71: 1-15.
Teller, P. (2001). Twilight of the perfect model. *Erkenntnis* 55: 393-415.
Temple, D. (1988). Discussion: The contrast theory of why-questions. *Philosophy of Science* 55: 141-151.
Thelen, E., Schöner, G. Scheier, C. & Smith, L. B. (2001). The dynamics of embodiment: A field theory of infant perseverative reaching. *Behavioral and Brain Sciences* 24: 1-86.
Thelen, E. & Smith, L. B. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge MA: MIT Press.
Time magazine U.S., (1945). *Medicine: Again, Trench Foot*. Monday Jan. 1. Retrieved 12-10-2012 from: http://www.time.com/time/magazine/article/0,9171,791832,00.html
Toon, A. (2012). *Models as Make-Believe: Imagination, Fiction, and Scientific Representation*. Basingstoke: Palgrave Macmillan.
Treiber, C. D., Salzer, M. C., Riegler, J., Edelman, N., Sugar, C., Breuss, M., Pichler, P., Cadiou, H., Saunders, M., Lythgoe, M., Shaw, J. & Keays, D. A. (2012). Clusters of iron-rich cells in the upper beak of pigeons are macrophages not magnetosensitive neurons. *Nature* 484, 367-370.
Tsukiura, T., Mano, Y., Sekiguchi, A., Yomogida, Y., Hoshi, K., Kambara, T., Takeuchi, H., Sugiura, M. & Kawashima, R. (2010). Dissociable roles of the anterior temporal regions in successful encoding of memory for person identity information. *Journal of Cognitive Neuroscience* 22: 2226-2237.

Tulving, E. (1972). Episodic and semantic memory. In: E. Tulving & W. Donaldson (eds.) *Organizations of Memory*. New York: Academic Press, pp. 381-403.
Van Bouwel, J. & Weber, E. (2008a). A pragmatist defense of non-relativistic explanatory pluralism in history and social science. *History and Theory* 47: 168-182.
Van Bouwel, J. & Weber, E. (2008b). De-ontologizing the debate on social explanations: a pragmatic approach based on epistemic interests. *Human Studies* 31: 423-442.
Vanderbeeken, R. & Weber, E. (2002). Dispositional explanations of behavior. *Behavior and Philosophy* 30: 43-59.
Van Fraassen, B. (1980). *The Scientific Image*. Oxford University Press.
Van Gelder, T. (1995). What might cognition be, if not computation? *Journal of Philosophy* 91: 345-381.
Van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences* 21: 615-665.
Van Gelder, T. & Port, R. F. (1995). It's about time: an overview of the dynamical approach to cognition. In: R. F. Port & T. van Gelder (eds.), *Minds as Motion*. Cambridge MA: MIT Press, pp. 1-43.
Walmsley, J. (2008). Explanation in dynamical cognitive science. *Minds and Machines* 18: 331-348.
Weber, E., Gervais, R. & Van Bouwel, J. (2013). The 'green cheese' and 'red herring' problems reconsidered. Epistemological vs. methodological tasks for philosophers of science. *Manuscript submitted for publication*.
Weber, E. & Van Bouwel, J. (2007). Assessing the explanatory power of causal explanations. In: J. Persson & P. Ylikoski (eds.), *Rethinking Explanation*. Dordrecht: Kluwer Academic Publishers, pp. 109-118.
Weber, E., Van Bouwel, J. & De Vreese (In press): *Scientific Explanation*. Springer.
Weber, E., Van Bouwel, J. & Vanderbeeken, R. (2005). Forms of causal explanation. *Foundations of Science* 10: 437-454.
Weber, E. & Vanderbeeken, R. (2005). The functions of intentional explanations of actions. *Behavior and Philosophy* 33: 1-16.
Wheeler, M. (2005). *Reconstructing the Cognitive World: The Next Step*. Cambridge MA: MIT Press.
Wheeler, M. & Clark, A. (1999). Genic representation: reconciling content and causal complexity. *British Journal for the Philosophy of Science* 50: 103-135.
Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., Nakayama, K. & Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences of the Unites States of America,* 107 (11): 5238-5241.
Wilson, R. A. (2004). *Boundaries of the Mind: The Individual and the Fragile Sciences*. New York: Cambridge University Press.
Woodward, J. (2001). Law and explanation in biology: invariance is the kind of stability that matters. *Philosophy of Science* 68: 1-20.
Woodward, J. (2002). What is a mechanism? A counterfactual account. *Philosophy of Science* 69 (supplement): S366-S377.
Woodward, J. (2003). *Making Things Happen*. Oxford: Oxford University Press.
Wright, C. D. (2012). *Mechanistic explanation without the ontic conception*. European Journal for Philosophy of Science 2: 375-394.
Yin, R. K. (1969). Looking at upside-down faces. Journal of Experimental Psychology 81: 141-145.
Zednik, C. (2011). The nature of dynamical explanation. *Philosophy of Science* 78: 238-263.
Zhu, Q., Song, Y. Hu, S. Li, X., Tian, M., Zhen, Z., Dong, Q., Kanwisher, N. & Liu, J. (2010). Heritability of the specific cognitive ability of face perception. *Current Biology,* 20 (2): 137-142.