THE UNIVERSITY of EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Concept Learning Challenged

Richard Stöckle-Schobel

PhD (Philosophy)
The University of Edinburgh
2013

Abstract

In my thesis, I argue that the philosophical and psychological study of concept-learning mechanisms has failed to take the diversity of learning mechanisms into account, and that consequently researchers should embrace a new way of thinking about concept learning: 'concept learning' as a class of psychological mechanisms is not a natural kind lending itself to unified study and should be eliminated. To arrive at this, I discuss several concept-learning models that attempt to overcome Jerry Fodor's challenge and base my judgment on the plurality of feasible concept-learning mechanisms and on criteria for theoretical notions from the philosophy of science.

Chapter 1 serves as an introduction to the topic 'concept learning' and highlights its importance as a research topic in the study of the mind. I argue that a mechanistic understanding of the shape of concept learning is best suited to explain the phenomena, in line with the recent resurgence of mechanism-based explanation in the philosophy of mind. As the main challenge to the idea that concepts can be learnt, I proceed to set up Fodor's challenge for concept learning in Chapter 2. This challenge is the idea that concepts cannot be learnt given the logically possible mechanisms of concept learning. I lay out the argumentative structure and background assumptions that support Fodor's argument, and propose to scrutinise his empirically based premise most closely in my thesis: this empirically based premise is that the only possible mechanism of concept learning is the process of forming and testing hypotheses.

As replies to Fodor's challenge, I discuss Perceptual Learning (R. Goldstone), Perceptual Meaning Analysis (J. Mandler), Quinean Bootstrapping (S. Carey), pattern-governed learning (W. Sellars), joint-attentional learning (M. Tomasello), and the Syndrome-Based Sustaining Mechanism Model (E. Margolis and S. Laurence). I argue that almost every mechanism I discuss has some leverage against Fodors argument, suggesting that there may be a wide variety of non-hypothesis-based concept-learning mechanisms.

The final chapter of my thesis, Chapter 7, takes a step back and reviews the fate of the notion of concept learning in light of the diverse set of learning mechanisms brought up in my thesis. My first and main worry is that it is questionable whether the previously discussed mechanisms of concept learning share many scientifically relevant properties that would justify seeing them as instances of the natural kind 'concept learning mechanism'. I argue that the substantiation of this worry would necessitate the elimination of 'concept learning' and 'concept-learning mechanism' as terms of the cognitive sciences. The chapter lays out the argumentative structure on which Concept Learning Eliminativism (CLE) rests, along with a discussion of questions about natural kinds and pragmatics in theory construction. This is inspired by Edouard Machery's argument for the elimination of 'concept', but independent of Machery's own project.

With this in place, I go on to give a conclusive argument that supports CLE, based on the claims that ‘concept learning’ is not a natural kind and that there are pragmatic advantages to eliminating ‘concept learning’. In this final chapter, I also raise pragmatic considerations that support the argument for CLE, and propose new research directions that could profit from the eliminativist position.

Contents

Declaration v

Note on publications vi

Acknowledgements vii

1 Introduction 1

1.1 The tale of the aardvark 2

1.2 Challenging concept learning 2

1.3 Mapping the territory 3

1.3.1 What is a concept? 3

1.3.2 Desiderata for a theory of concepts 4

1.3.3 What is concept learning? 6

1.3.4 Why should we look for mechanisms? 7

1.4 The structure of the thesis 7

2 Fodor's Challenge for theories of concept learning 10

2.1 Introduction 11

2.2 Fodor's paradox 11

2.3 Innateness in Empiricist and Rationalist theories 13

2.4 A more detailed look at the steps of Fodor's argument for the paradox . 15

2.4.1 (4) The Hypothesis-Formation-and-Confirmation Model 15

2.4.2 (7)+(8) Spelling out the paradox 16

2.4.3 (3) The representational theory of mind (RTM) 16

2.4.4 (9) The anti-learning conclusion 19

2.5 The doorknob/DOORKNOB problem 21

2.5.1 The metaphysical solution 22

2.5.2 The neurological solution 23

2.5.3 The upshot of Fodor's position 26

2.6 Fodor's Challenge 27

3 Developmental accounts of concept learning 29

3.1 The challenge for developmental psychology 30

3.2 The 'Evidence for Prelinguistic Concept Learning' hypothesis 30

3.2.1 The dominant research paradigms in early infancy 31

3.2.2 The argument against EPCL 34

3.3 Jean Mandler 40

3.3.1 Perceptual Meaning Analysis 42

3.3.2 Empirical evidence for PMA 43

3.3.3 Problems for PMA 44

3.4 Susan Carey 48

3.4.1 Foundations of Quinian Bootstrapping 49

3.4.2 Analysis of Quinian Bootstrapping 51

3.4.3 Evidence for Quinian Bootstrapping 54

3.4.4 Problems for Quinian Bootstrapping 54

3.5 Summary 57

4 Perceptual Learning as a concept-learning mechanism 58

4.1 Perceptual Learning and feature-based approaches to concepts 59

4.2 Fixed-feature approaches and Fodor's paradox 59

4.3 Perceptual Learning as a reply to Fodor's Challenge 61

4.3.1 Unitisation 64

4.3.2 Differentiation 65

4.4 Concept learning or conceptual change? 67

4.4.1 Do unitisation and differentiation operate on concepts? 67

4.4.2 Intermission: Kinds of learning 68

4.4.3 Features as concepts – a related issue 70

4.4.4 Landy and Goldstone's reply: the Pask device 71

4.4.5 Intermediate conclusion 72

4.5 The notions 'feature' and 'stimulus' 72

4.5.1 The Compatibility criterion 72

4.5.2 The proximity of stimuli 73

4.5.3 Feature detectors and maturation 75

4.5.4 Physical information 75

4.6 Perspectives for an amendment of the Perceptual Learning approach . . 76

4.6.1 Three kinds of percept-concept interactions in Perceptual Learning 76

4.6.2 Blurring a boundary: brute-causal and rational acquisition . . . 78

4.7 Conclusion 81

5 Sellarsian concept learning 83

5.1 Introduction 84

5.2 The Augustinian Conception 84

5.3 Sellars's argument against AC – a reconstruction 87
5.3.1 The Awareness challenge 88
5.3.2 The Holist challenge 89
5.3.3 The Relational challenge 90
5.4 Sellarsian concept learning 92
5.4.1 Logical spaces 92
5.4.2 Learning without 'knowing the rules' 95
5.4.3 The advantages and the problem of Sellars-type learning 97
5.5 Gauker's view: imagistic cognition before conceptual thought 99
5.5.1 Categorisation tasks in early infancy 100
5.5.2 Are similarity judgments conceptual? 101
5.6 Developmental evidence for prelinguistic concepts 103
5.6.1 Person, agent, gaze-following 103
5.6.2 Social referencing 104
5.6.3 Pointing 105
5.6.4 A cut-off point for speaking of concepts? 106
5.7 A second kind of social concept learning mechanism: Tomasello's Social-Pragmatic Theory of Word Learning 107
5.7.1 The Social-Pragmatic Theory of Word Learning 108
5.7.2 Examining the fit of SPT and Sellars's model 109
5.7.3 Standing up against Fodor's Challenge 109
5.8 Conclusion 110

6 Learning concepts with sustaining mechanisms 111
6.1 Introduction 112
6.2 Margolis and Laurence's response to Fodor 113
6.2.1 The simple Fodor puzzle 113
6.2.2 Arguments for complex concept learning 113
6.2.3 The argument for primitive concept learning 114
6.2.4 The syndrome-based sustaining mechanism model 117
6.2.5 Things to like about the sustaining mechanism model 121
6.2.6 Problems for the SBSM 123
6.3 Criteria for concept learning 126
6.4 Between brute-causal and rational acquisition 130
6.4.1 The HF model and SBSM learning 131
6.4.2 The Pask device's learning 131
6.4.3 Could triggering qualify as learning by this measure? 134
6.4.4 Lessons from applying the learning criteria 135
6.5 Conclusion: Defending the rationality of concept acquisition 136

7 Concept learning eliminated 138
7.1 The state of play 139
7.2 The claim 139
7.3 The Horizontal Argument for Scientific Eliminativism 140
7.3.1 Types of eliminativist arguments 141
7.3.2 Natural kinds 144
7.3.3 Pragmatic grounds 146
7.4 The argument for Concept Learning Eliminativism 147
7.4.1 Against CL as a natural kind 148
7.4.2 Pragmatic reasons for CLE 151
7.5 Problems with CLE 155
7.5.1 Concept Learning Pluralism 155
7.5.2 Mechanisms as natural kinds 156
7.5.3 Overgeneralisation 157
7.5.4 A primary kind of CLM 158
7.6 The consequences for concept learning 158
Conclusion 161
Bibliography 164

Declaration

I declare that this thesis is my own work, that it has been composed by myself, and that this thesis has not been submitted for any other professional degree or qualification.

Richard Stöckle-Schobel Edinburgh, 30.08.2013

Note on publications

A shortened version of chapter 4 has been published as:

Stöckle-Schobel, Richard (2012). Perceptual learning and feature-based approaches to concepts – a critical discussion. Frontiers in Psychology. 93(3).

Acknowledgements

In the course of working on my doctoral thesis, lots of people have helped me and have been of importance to me. I cannot name them all, even though I thank all of them. Some of them however deserve special mention.

Thank you to my thesis supervisors for their tireless support and help throughout the past years: Andy Clark, Julian Kiverstein, Tillmann Vierkant, and Mark Sprevak all challenged me and gave me philosophical inspiration. Andy and Julian have introduced me to many new ideas and always had exciting literature recommendations at their fingertips during the first two years of my studies. Till has been a constant and thought-provoking interlocutor and can't be thanked enough for his help throughout my time in Edinburgh. Finally, Mark has taken over primary supervisor duties in 2011, and has guided me through the whole process of completing this work. Many ideas that are central to this thesis emerged during our meetings, and I couldn't have wished for a more knowledgeable and kind supervisor.

I am deeply indebted to Michael Tomasello, who granted me a four-month research visit at the Max-Planck-Institute for Evolutionary Anthropology in Leipzig, and a generous scholarship for this visit. It was a great experience to work at his institute and to share and discuss ideas with him. Equally heartfelt thanks to Richard Moore, who worked most closely with me in Leipzig, and whose patience and generosity brought about the beginning of my PhD thesis in its current form. Even after I left Leipzig, he continued to share advice and comments on my work, for which I am especially grateful.

I want to thank the reviewers and journal editors for their comments and support leading up to the publication of my Frontiers in Psychology paper.

Many fellow students in Edinburgh and elsewhere have discussed their work, and on occasions even my work, with me, for which I am most grateful. Special thanks to the members of the Mind and Cognition discussion group for great debates and good times.

Thank you to the School of Philosophy, Psychology, and Language Sciences, who repeatedly granted me travel funds for presenting my work at academic conferences.

Special thanks to Katie Keltie, Toni Noble, and Lynsey Buchanan for their administrative help.

A big Thank You to my parents, Franziska and Franz, who supported me throughout my time in Edinburgh, and who made it all possible.

My deepest gratitude and love to my wife Catherine, who helped me through all of this, and to whom I dedicate this work.

Chapter 1

Introduction

1.1 The tale of the aardvark

For most of my life, I didn't know what an aardvark is. As far as I could tell, aardvarks could have been tropical birds, or 'aardvark' could have been a Norwegian word for nightmare. I found out about the existence of the word by reading philosophical articles on concepts, and learnt that the aardvark is a kind of African mammal, similar to the anteater, by browsing the Internet for information about aardvarks. This is the short and unremarkable story of how I learnt the concept AARDVARK. It is very clear that it isn't the full psychological story of how I learn certain concepts, and much less is it a philosophical story of how new concepts are learnt. Before I can begin to engage such complete accounts of this part of human cognition, I want to show what kinds of stories one has to tell to give a full philosophical or psychological account of concept learning.

1.2 Challenging concept learning

The story I want to tell has a curious conclusion: as it turns out, many psychological mechanisms that are quite different in terms of their components all have a claim to be concept-learning mechanisms. More curiously, many of them are good exemplars of the kind, despite their differences. This leads me to the conclusion that they cannot all be or do the same thing – that they cannot all be instances of a general class of things (a natural kind) that are concept-learning mechanisms. Since, by this conclusion, concept-learning mechanisms are not a natural kind, I propose to eliminate the notion 'concept-learning mechanism' from the cognitive sciences.

To establish this conclusion, I develop the following points in the chapters that lead up to it. First, the question whether concepts can be learnt at all is one that philosophers and psychologists have debated for decades, and still debate today. To get a grip on it, I discuss an influential position against the possibility of concept learning – Jerry Fodor's anti-learning stance – and show that Fodor's argument against concept learning leaves proponents of learning in an uncomfortable position. I further argue that Fodor's own proposal for concept acquisition without learning faces serious problems. These two aspects make the search for alternative proposals pressing. I further argue, through the investigation and discussion of several proposed concept-learning mechanisms from developmental psychology, cognitive psychology, and philosophy of mind, that there are numerous proposals that can overcome Fodor's argument. They all do it by different means, though. This leads me to the conclusion I started with – the differences between the available, successful, concept-learning mechanisms weigh so heavily that there are legitimate reasons to abandon grouping them together under the heading 'concept-learning mechanism'.

1.3 Mapping the territory

Concepts are at the core of philosophical and psychological theories of human cognition. As a convention, when I talk about a concept, for instance the concept of meteorites, I will use small caps – the concept METEORITE. When I want to talk about the name of the concept – the word that stands for it – I will put it into single quotes – 'meteorite' in English. When I want to talk about meteorites themselves, I will use the word without special signs (except when I want to emphasise it, in which case I will italicise it).

Answering the following three questions will put me in a position of giving a concise overview of the main argument of my thesis, including the necessary steps to arrive at the point of making it. These questions are:

1. What is a concept?
2. What is concept learning?
3. Why should we look for mechanisms?

1.3.1 What is a concept?

Many philosophers agree on a very general characterisation of concepts, such as the one Jesse Prinz (2002) puts forward:

> (...) concepts are constituents of thoughts. (Prinz, 2002, p. 2)

To a curious reader, this might not be an enlightening description by itself; for example, it doesn't spell out why they are a particularly interesting class of constituents of thought. There might be much more important, mysterious, or else interesting constituents of thoughts. Assuming that concepts indeed are important, in some way mysterious, and interesting, we need a more specific description. For instance, it would seem natural to pay special attention to *propositional* thought – what would the constituents of a thought like I EXPECTED GRATITUDE be? The key to analysis here is to divide the thought into its individual parts, and regard these as concepts: I, EXPECTED, and GRATITUDE.

For Jerry Fodor (1975), concepts are not just any kind of constituents of thoughts, as they appear to have some features that make them look distinctively like words in public languages. Fodor proposes to regard them as the elements of a special kind of language – the Language of Thought. The Language of Thought is supposed to be a set of atomic symbols that we use to form thoughts. These atomic symbols are unlike words in a natural language because they cannot be divided into smaller units of meaning. However, for others, like Christopher Gauker (2011), concepts are best understood as actual words. This means that a person without a public language would not possess any concepts. Gauker argues that conceptual thought is only one of several important aspects of human cognition; for example, thinking in images – imagistic cognition – has great importance for Gauker.

There is yet another interesting way philosophers have characterised concepts:

> Concepts, the old-fashioned logic-books tell us, are presupposed to, and exercised in, acts of judgment. (Geach, 1957, p. 11)

On this view, concepts are not necessarily like words. Indeed, Geach goes on to speak of concepts as abilities (cf. Geach, 1957, p. 13). A concept could be a certain cognitive capacity that can be put to a variety of uses, not just a verbal judgment: a pedestrian could exercise her concept SAFE PASSAGE by looking right and left before crossing a street, finding it to apply just in case there is no traffic in her direct vicinity.

A third view on concepts can be found in cognitive psychology as Machery (2009) argues. On this view, concepts are

> 'bodies of knowledge used by default in the processes underlying most higher cognitive competences.' (Machery, 2009, p. 239)

This way of talking about concepts has many advantages. Most notably, it classifies concepts by the function they play for psychologists. Rather than just raising the issue of conscious and propositional thought, it refers to 'higher cognitive competences', which comprise a wide range of capacities, like categorisation and reasoning, which are not at the centre of the philosophical study of concepts.

In my thesis, I will not commit to any one of the proposed theories or views about what concepts are. By keeping the notion neutral, and only requiring the concepts I speak about to conform roughly to Prinz's very general characterisation that I have given above, I can take a diverse range of proposed concept-learning mechanisms into account. This will aid in further understanding the learning process. I will, however, exclude some options in the course of my work: for example, I will exclude Gauker's view that concepts are words by arguing, in Chapter 5, that there are good reasons to regard prelinguistic infants as concept-users.

1.3.2 Desiderata for a theory of concepts

Prinz (2002, ch. 1) offers a list of desiderata for theories of concepts, which he proceeds to test on the main theories in the philosophical and psychological literature. He proposes that "these desiderata can serve as a litmus test for a theory of concepts" (Prinz, 2002, p. 3), which means that they can serve as the measure of the quality of theories of concepts. I want to list them and give a brief comment on each, to further elucidate what is at stake philosophically and psychologically when we talk about concepts:

1. Scope: A theory must explain how we can reason about, and use abstract concepts just as well as perceptual concepts, and unobservable and intangible concepts just as everyday concepts, and natural kind concepts.

2. Intentional Content: A theory of concepts must explain how the contents of concepts can refer to things in the world and to other extramental referents. That means that a theory of concepts must be able to explain the intentionality of concepts.

3. Cognitive Content: A theory of concepts must also explain the relations or lack of relations between concepts in our own introspection or cognitive awareness. If we fail to identify Tully, the politician, with Tully, the orator, then we need an explanation in terms of our concepts for that.

4. Acquisition: A theory of concepts must be able to explain how concepts can be acquired. I will go into greater detail on this desideratum in the next section.

5. Categorisation: A theory of concepts must explain the human ability to categorise instances of concepts into the right kinds of categories. Categorisation is central for object identification, object recognition, and abstract thought, since it allows a thinker to discern the relevant features, and the (good) instances of a concept.

6. Compositionality: A theory of concepts must account for the compositionality of thought. This means that it has to explain how it is possible to combine concepts into thoughts or into more elaborate concepts. Crucially, compositionality is needed for an explanation of the productivity and systematicity of thought, which many philosophers regard as central features of human cognition.

7. Publicity: A theory of concepts must be able to explain how people can share concepts, or how they can use the same concepts to denote the same things.

Prinz considers an eighth desideratum – a language desideratum:

8. Language: A theory of concepts must be able to explain how concepts are related to language.

He argues against the inclusion of this desideratum because he thinks that a cogniser can have concepts without having language. Thus, language is not essential to concept possession, and shouldn't be included in a list of things that a theory of concepts should be able to account for. I agree with this assessment, and will therefore not regard the language desideratum as central to understanding concepts.

1.3.3 What is concept learning?

As we have seen, one of Prinz's desiderata, which only gets a cursory treatment in his book, is 'Acquisition':

> (...) a theory must ultimately support a plausible explanation of how concepts are acquired. (Prinz, 2002, p. 8)

Prinz goes on to separate the demand for a plausible explanation of concept acquisition into an ontogenetic and a phylogenetic strand. We want to know how we acquire concepts over the course of a lifetime, which is the question of ontogeny. Furthermore, we want to know how some concepts, or some conceptual capacities, or the tools for acquiring some concepts rather than others, are transmitted over evolutionary time scales – this is the question of phylogeny. In my thesis, I will focus on the ontogenetic question.

Among the philosophers who have worked on concept learning, Jerry Fodor occupies an important spot; he has extensively discussed the matter throughout his career and is among the most outspoken defenders of the Rationalist tradition in contemporary philosophy of mind. Indeed, Fodor draws on ideas from Plato, St. Augustine, Descartes and Leibniz, among others, in his work on concepts and on mental representation. His work has been influential in other areas of the cognitive sciences, and he has collaborated with linguists and psychologists. All of this makes him an important interlocutor on questions of the nature of concepts, so I will put his views at the centre of my thesis. In the next chapter, I will go into detail on Fodor's views regarding concept learning, but for now I want to draw attention to an important distinction that he draws between concept learning and concept acquisition. He holds the view that learning is a special kind of acquisition – to wit, it is the kind of acquisition that is guided by rational and causal processes (cf. Fodor, 1981, p. 273). By talking about rational processes, Fodor means to draw attention to the cognitive nature of the process. Learning has to come about through cognitive states that are related to some kinds of evidence or instance of a concept that is about to be learnt. As we shall see soon, Fodor regards this relation as that between a hypothesis and a piece of evidence (dis-)confirming it. The hypothesis, or in any case the evidence-related state of mind, will be central in acquiring a new concept. Fodor puts the central point as follows:

> The theory that concepts are acquired by learning does explain why concepts very often fit the world: it's because they are very often learned *from* the world by a generally reliable inferential procedure. (Fodor, 2008, p. 149)

So, concept learning is one central kind of process for acquiring a concept. In later chapters, I will discuss more thoroughly the distinction between acquisition and learning.

1.3.4 Why should we look for mechanisms?

There is a growing literature in the philosophy of science that investigates the nature of mechanistic explanations (Machamer et al., 2000; Craver, 2006, 2009; Bechtel, 2009; Bechtel and Abrahamsen, 2010, among many others). In contrast to other sciences, such as physics, special sciences like psychology don't tend to use laws in explaining psychological phenomena (cf. Bechtel and Wright, 2009, p. 113). Rather, psychologists rely on mechanistic explanations for psychological processes. Bechtel and Wright (2009) define 'mechanism' as follows:

> A mechanism is simply a composite system organized in such a way that the coordinated operations of the component parts constitute the mechanistic activity identified with the explanandum. (Bechtel and Wright, 2009, p. 119)

An advantage of the mechanistic account of psychological explanation is that it attempts to find the psychological components that are involved in cognitive activity, and that it can provide explanations of the causal relations between these component parts.

I will assume that mechanistic explanation is best suited for exploring concept learning. It should be apt to discover the psychological components, biases, or structures that are needed to create new cognitive tools or constituents for thought. Given that concept learning, especially from perceptual experience, is bound to be a complex process involving a variety of components (perception, memory, prior concepts, language capacities, to name just a few), an explanation that captures the coordination between these components as well as the causal links between them is needed. Mechanistic explanation is the paradigm for this kind of psychological research.

Now that I have provided answers to the three main questions at the beginning of this chapter, I will begin my investigation. To help the reader in getting an overview of the thesis, I give below a chapter-by-chapter summary of the main arguments.

1.4 The structure of the thesis

In the following, I argue that the philosophical and psychological study of concept-learning mechanisms has failed to take the diversity of learning mechanisms into account, and consequently that researchers should embrace a new way of thinking about concept learning: 'concept learning' as a class of psychological mechanisms is not a natural kind lending itself to unified study and should be eliminated. To arrive at this, I discuss several concept-learning models that attempt to overcome Jerry Fodor's challenge (Fodor, 1975, 1981, 2008). I base my judgment on the plurality of feasible concept-learning mechanisms and on criteria for theoretical notions from the philosophy of science.

As the main challenge to the idea that concepts can be learnt, I set up Fodor's concept learning paradox in Chapter 2. This challenge from the paradox is that concepts cannot be learnt given the possible mechanisms of concept learning. I lay out the argumentative structure and background assumptions that support Fodor's argument. I scrutinise Fodor's empirically based premise most closely in my thesis. This empirically based premise is that the only possible mechanism of concept learning is the process of forming and testing hypotheses (HF model). I argue that Fodor's paradox poses a serious problem, and that Fodor's own attempts to solve it, by postulating concept-acquisition mechanisms that don't count as concept learning, fail. This results in what I call Fodor's challenge: to overcome Fodor's paradox, one must offer a concept-learning mechanism cannot be interpreted as an instance of the HF model.

In Chapter 3, I set up the first two contenders for concept-learning mechanisms from Susan Carey and Jean Mandler. Both offer a theory of concept learning that is grounded in developmental psychology. Jean Mandler's Perceptual Meaning Analysis places heavy weight on the notion that perceptual content can be transformed into conceptual content through analysis (Mandler, 2004). Susan Carey, on the other hand, regards concept learning as a bootstrapping process that is necessary because of the incommensurability of children's and adults' concepts (Carey, 2009). I critically discuss the role that developmental evidence from early ontogeny plays for these two theories, and argue a) that Mandler's theory is not supported by the evidence she gives, and b) that Carey escapes the difficulty insofar as her proposal relies on evidence from later childhood, and not on data from prelinguistic infancy.

In Chapter 4, I discuss Robert Goldstone's Perceptual Learning approach (Landy and Goldstone, 2005; Goldstone and Landy, 2010), which posits the mechanisms of unitisation and differentiation as the main agents in concept learning. Two problems keep me from accepting it as a successful answer to the challenge. First, Goldstone's Perceptual Learning seems to operate with existing concepts, which it then unites or differentiates; this would not count as concept learning by Fodor's standard, since there is strictly speaking no addition of *new* concepts to the cognitive system. Second, I raise doubts about the possibility of Perceptual Learning fulfilling the compatibility criterion, which requires new concepts to be compatible to the old ones; it must be possible to combine previously available concepts and new, perceptual concepts in forming new thoughts. Goldstone doesn't address the constraints upon the possibility of combining concepts from perceptual origins with amodal and abstract concepts. However, exploring the possibilities of Perceptual Learning raises two ways in which alternative conceptions of concept learning, such as Goldstone's, might be able to replace Fodor's HF model. These are, first, the possibility of using perceptual experience to change one's conceptual structure, and second, the proposal to include 'brute-causal' mechanisms as inputs for learning episodes. I re-examine and develop the latter idea in

Chapter 6.

In Chapter 5, I focus on Wilfrid Sellars's theory of concept learning as pattern-governed behaviour (Sellars, 1969, 1974). While Sellars's rejection of the Augustinian Conception and of Fodor's concept possession condition offer him the possibility to centre the explanation of concept learning on its social aspects, it also has a major lacuna that a Fodorian might be tempted to fill with her own hypothesis: Sellars's model cannot explain the development of prelinguistic representational capacities. To salvage Sellars's position, I propose a way of joining his ideas with a developmental theory of joint-attentional learning based on the work of Michael Tomasello (2003) (see also Tomasello, 1999; Tomasello et al., 2005). With that, we get a new aspect of concept learning: social concept learning transcends Fodor's challenge insofar as it achieves concept learning without the HF model.

In Chapter 6, I discuss the Syndrome-Based Sustaining Mechanisms model developed by Stephen Laurence and Eric Margolis. Since this model conforms closely to Fodor's own demands for a theory of concepts and of representation, Fodor struggles to counter this model with strong opposition. In the chapter, I find good reasons to believe that Margolis and Laurence overcome Fodor's challenge. I show that the objections to their model aren't successful. The discussion of concept-learning mechanisms isn't over though: since Margolis and Laurence (2011) offer a set of criteria for concept learning which is met by several models, I look more at the possibility of acknowledging a wide range of concept-learning mechanisms between just-not-brute-causal acquisition and the HF model. The result is a preliminary endorsement of pluralism about concept-learning mechanisms.

The final chapter of my thesis, Chapter 7, reviews the fate of the notion of concept learning in light of the diverse set of learning mechanisms brought up in my thesis. My main concern is whether the previously discussed mechanisms of concept learning share many scientifically relevant properties that would justify seeing them as instances of the general natural kind 'concept-learning mechanism'. I argue that this worry would necessitate the elimination of 'concept learning' and 'concept-learning mechanism' as terms from the cognitive sciences. The chapter lays out the argumentative structure on which Concept Learning Eliminativism (CLE) rests, along with a discussion of questions about natural kinds and pragmatic considerations in theory construction. This work is inspired by Edouard Machery (2009)'s argument for the elimination of 'concept', but it is independent of Machery's project. With this in place, I go on to give an argument that supports CLE based on the claims that 'concept learning' is not a natural kind and that there are pragmatic advantages to eliminating 'concept learning'. In this final chapter, I also propose new research directions that could profit from the eliminativist position.

Chapter 2

Fodor's Challenge for theories of concept learning

2.1 Introduction

Some philosophers, like Socrates in Plato's Meno and Phaedo and, most prominently in recent debates, Jerry Fodor, thoroughly question the possibility of genuine concept learning from experiences and deny a content-generating link between perception and conception. Fodor (1975) has formulated a concise argument for this claim, which was reformulated in Fodor (1981), slightly changed in Fodor (1998) and had its latest instalment and strongest claim in Fodor (2008). In the following, I will present Fodor's argument, link his claims to their historical background in the Rationalism/Empiricism debate of the Early Modern period, and place his argument within the framework of philosophical commitments he operates in. Based on this presentation of Fodor's view on concept learning, I will introduce the doorknob/DOORKNOB problem as the main problem his anti-learning stance has to overcome, and argue that Fodor fails to give a satisfying account of concept acquisition without concept learning. To conclude this chapter, I will formulate my argumentative strategy for addressing Fodor's challenge to theories of concept learning.

I will base my analysis of what might be called 'the Learning Paradox' on its discussion in Fodor's *LOT2. The language of thought revisited* (Fodor, 2008). Aside from being Fodor's current position, and therefore deserving the closest consideration from among his anti-learning arguments, there are additional pragmatic reasons for doing so. The first one is that the latest argument is only based on one empirical premise, and leaves out another one which weakened Fodor's argument. The second reason is that I want to present the argument without presupposing the rest of his philosophy of mind. While there are good reasons for connecting the learning paradox with Fodor's views on the necessity of a Language of Thought, the dominance of the Computational Theory of Mind, and his position on the feasibility of causal theories of reference, such a discussion would exceed the limits of my investigation. The *LOT2* argument is best suited for the treatment which I plan to give in what follows, as it focuses on the essential points regarding concept acquisition.

2.2 Fodor's paradox

Although Fodor doesn't formalise his argument against concept learning, the overall structure can be put as follows. Each of these short points will be elucidated and discussed in this chapter.

1. (Conceptual clarification 1): Learning mechanisms are "rational-causal processes" (Fodor, 1981, p. 273). Being a rational process, learning is mediated by psychological states, such as beliefs. Being a causal process, learning proceeds from causes (such as experiences of an object) to effects (learnt concepts pertaining to that object).

2. (Conceptual clarification 2): What is not learnt is innate or acquired in some other non-rational way.

3. (Supposedly self-evident truth 1): "A sufficient condition for having the concept C is: being able to think about something *as (a)* C (being able to bring the property C before the mind as such, as I sometimes put it)." (Fodor, 2008, p.138) This means that the main act of concept-use is using the concept in forming beliefs (or other types of thoughts) – as contrasted with using the concept to categorise new sensory experiences, or to act upon a thing in the world. Thinking is prior to perceiving and acting in the order of concept use. This premise stems from Fodor's 'Cartesian Rationalism' (cf. Fodor, 2008, p. 8ff.).

4. (Empirical premise 1): The only available, empirically tested model for learning is the following: learning the concept C consists in forming hypotheses about C and testing them against the available evidence. Thus, learning is a process of inductive inference.

5. (Conditional premise 1): If the hypothesis-formation model (HF model) is the only available model for learning concepts, then all instances of concept learning are based on the HF model.

6. (Modus Ponens, 4, 5): All instances of concept learning are based on the HF model.

7. (Exemplification of 6): Forming a hypothesis about the concept C requires bringing the property expressed by C before one's mind. One needs to think about a piece of evidence 'x' as (a) C to (dis-)confirm hypotheses about C. To learn which things are green, one must judge something to be (or not be) green. This act of judging is a mental activity for which one needs to be able to think about green things as green things.

8. (4+7): What is already used for hypothesis formation is not learnt in the application (confirmation or disconfirmation) of the hypothesis. C was already available to form the hypothesis, thus C was not learnt.

9. (2+8): All concepts are innate or acquired in some other non-rational way.

The first four premises make up the body of background assumptions that are required for the argument's conclusion, (9), to go through. (7) and (8), on the other hand, are statements that are supposed to highlight the consequences of the premises, and lead to (9). So, one can see that the argument is based on different types of assumptions: (1) and (2) are important as terminological constraints, and have philosophical as well as psychological ramifications, which I will discuss below, and which will shape the remainder of this thesis. (3), on the other hand, is something like a Rationalist doctrine, rooting Fodor's argument in the philosophical tradition of Descartes and Leibniz. (4) serves double duties for the argument, as the only empirical premise and as the grounds for an 'Inference to the Best Explanation' (IBE) argument. I will devote a considerable amount of discussion to this premise and the underlying assumptions in this thesis, and will argue that an argument against Fodor's view is most fruitfully targeted at this premise.

The aim of Fodor's argument is to show that the concept learning has no philosophical or psychological warrant. If the HF model is the only scientifically plausible mechanism of learning, and if it works in the way Fodor takes it to work, then we face a paradox. Learning the concept x already requires the possession of x. But if x is already available to the learner, then she doesn't actually acquire anything new by hypothesising about x. Thus, the notion of concept learning is empty, and we need to look for another way to explain how humans acquire the concepts that they have. By premise (1), it cannot be a process that is in some way mediated by psychological states, like beliefs or hypotheses. But at the same time, a process that doesn't rely on the formation of beliefs of some kind also doesn't seem like a possible candidate for concept learning, if we take Fodor's philosophical views seriously: without the involvement of beliefs, Fodor wouldn't call the process 'rational', so we wouldn't have a rational-causal process before us. It would rather be something like an unconscious bit of cognitive processing that our kinds of brains just do, and we might object to calling this kind of processing 'learning'.

This state of affairs is what I would like to call *Fodor's paradox.*

2.3 Innateness in Empiricist and Rationalist theories

Fodor's conclusion does serious damage to Empiricist accounts of conceptual development, which typically build upon the idea that we acquire most of our conceptual repertoire by learning from our experiences. Remember the Empiricist doctrine that nothing is in the mind that was not first in the senses, as held by John Locke (see e.g. Locke, 2008, book II, chapter 1, §2).[1] The Empiricist program builds upon premise (1) above: learning about the world is something that we're actively engaged in, thinking about what we experience and extracting information from these experiences. On this view, the world is a resource of information and knowledge. In experiencing it, this information enters our cognitive system – we form a mental state that might be called "an experience of x." This experience is then a part of our mental life and we can use it to form new thoughts about this x and interlink it with other experiential knowledge. Notice that this way of talking is deeply rooted in the representational theory of mind in the form held in Early Modern times. A more careful, or at least more scientifically informed way of talking about representation is surely necessary to advance the present

[1] All references to historical texts such as Locke's are going to be to modern reprints, with the original publication date added in the bibliography.

discussion, and will be introduced below.

In Early Modern thought, Empiricism's main opponent was Rationalism. Whereas the strongest reading of Empiricism is the 'tabula rasa' idea of the mind being a 'blank slate' at birth, Rationalists typically opt for Leibniz's 'marble block' metaphor (Leibniz, 1996): the veins in a marble block predispose it to be shaped into certain forms; the mind, in analogy, has certain predispositions, or lines, along which it will be shaped once it comes into existence. Experience at best guides this shaping, but it cannot put anything into the mind that hadn't potentially been there before. Seeing himself as a Rationalist in the tradition of Descartes, Fodor is a proponent of innate concepts. This means that conceptual contents can't be derived or extracted from experience – rather, we need the conceptual contents to be able to accurately perceive and experience the world (cf. Fodor, 2008, p. 12).

In Fodor (1981), he observes that the main controversy between the Empiricist and the Rationalist is not about whether there are innate concepts, but rather about how big the basis of innate concepts is. He analysed that both views have an identical stance on 'simple ideas', or 'primitive sensory concepts' – concepts pertaining to the basic properties observable for the human senses:

> On all standard theories the sensorium [i.e. the ensemble of human senses] is an innately specified function from stimuli onto primitive concepts. (Fodor, 1981, p. 276)

If we hadn't the sensory organs we have, we wouldn't have the primitive sensory concepts that we have, hence we should treat these primitive concepts as innate, or so the reasoning goes in Locke, and in Fodor. So, how big is the innate basis of concepts? For classical Empiricists like Locke, this basis is very small (only the 'primitive sensory concepts'), whereas for classical Rationalists like Descartes, most, if not all, concepts are innate. Based on their acceptance of a potentially large class of innate concepts, Fodor (1981) argues that both Empiricists and Rationalists are basically Nativists, albeit to different degrees. If we suppose this is the case, is there an actual difference between the two schools of thought left? Fodor wants to argue that the difference lies in the status of lexical concepts (cf. Fodor, 1981, p. 278f.) – concepts that are expressed by a single word in the English language.

For Fodor, again following the Early Modern tradition, concepts are much like lexical items, and these items are either unstructured or structured. Phrasal concepts like THE ONLY LIVING BOY IN NEW YORK are of course structured, and therefore also complex. Lexical concepts are more complicated, however, as some of them can be seen as structured, while others are best thought of as unstructured. The choice depends on further philosophical commitments, in this case whether one adheres to Empiricism or Rationalism. A possibly structured concept, like BACHELOR, can be given a definition in terms of more primitive concepts, like UNMARRIED MAN (which is clearly a phrasal concept). An unstructured, or primitive, concept like GREEN can't be divided up into more primitive concepts.

Empiricists leave room for structured lexical concepts "because lexical concepts are normally constructs out of primitive concepts" (Fodor, 1981, p. 278). If lexical concepts were structured, then there would be some sense to using the HF model as a model for learning concepts, in the sense that it would explain how certain basic properties, taken together, form a meaningful unit: the concept TRIANGLE, although lexical, would be constructed from the defining set of more primitive concepts like THREE, SIDES, CORNERS, and a description of their spatial configuration, or something along those lines. For a Rationalist, however, the lexical concepts might as well be unstructured, since the mechanisms of concept acquisition are non-psychological anyway (more on this in section 2.5).

2.4 A more detailed look at the steps of Fodor's argument for the paradox

Having stated the raw structure of the Learning Paradox and the aim behind the argument, I want to proceed by looking closer at some of the argument's steps. Since it is central to the argument, and will play the most important role in the replies to Fodor which I discuss in my thesis, I want to start with the empirical premise and then touch upon the other premises along the way.

2.4.1 (4) The Hypothesis-Formation-and-Confirmation Model

Fodor conceives of learning as the building and testing of hypotheses (the Hypothesis-Formation-and-Confirmation model, hereafter also 'HF model'): a tentative statement about an object x is made, like a categorisation of x as an F, and then it is compared to the available evidence concerning Fs. In Fodor (1981), he gives an example of concept-learning studies in experimental psychology which he claims to be "a typical concept learning experiment" (Fodor, 1981, p. 266). In his example – which resembles the Wisconsin Card Sort Test, developed by Berg (1948), and the experiments in Bruner, Goodnow & Austin (1956) – the subject has to categorise a set of stimulus cards with coloured geometric figures into those which are 'flurg' and those which aren't (with the aim of learning the concept FLURG). In the example, Fodor defines 'flurg' as 'green or square'. Upon getting a new card, the subject has to state whether she judges the picture on the card to be a picture representing FLURG. The experimenter will then either confirm the hypothesis, say, because the object is green but not square, or disconfirm it, say, because it's a red circle. Of course, the experimenter will not give those reasons but only say "yes" or "no" upon hearing the subject's hypothesis. In this task, one usually starts by guessing, and only subsequently builds up a base of hypotheses, which get (dis-)confirmed. The hypotheses one builds have the form of biconditionals:

"Flurg(x) is true $\leftrightarrow$ F(x)" ('F' is a placeholder for an as yet unknown

> set of properties with, in this example, the final successful hypothesis being "Flurg(x) is true ↔ G(x) v S(x)" where 'G(x)' stands for 'x is green' and 'S(x)' stands for 'x is square')

Only over a certain amount of trials does one come to learn that 'flurg' means 'green or square'. The learning in this experiment consists in making the right category choice for a given number of consecutive trials. After reaching this quota, the experimenter judges that one has learnt the concept FLURG.

Fodor claims that this kind of experiment "informs practically all the psychological work on concept attainment which the Empiricist tradition has inspired." (Fodor, 1981, p. 267) Indeed, he assumes that all cognitive psychologists who adhere to the idea of concept learning have a model of inductive inference, like the one used in the Wisconsin Card Sort Test, in mind when endorsing the possibility of concept learning (cf. Fodor, 2008, p. 132).

2.4.2 (7)+(8) Spelling out the paradox

The argument is set to show that in order to categorise something as an F – here, as GREEN OR SQUARE – one already needs to have the concept F, or the concepts out of which F is constructed – in the present case the constituting concepts of GREEN, SQUARE, and OR. But if learning about F is conditional upon having the concept F, no concept learning can take place. All that happens is labelling of old concepts in new ways.

> Now, according to HF, the process by which one learns C must include the inductive evaluation of some such hypothesis as 'The C things are the ones that are green or triangular'. But the inductive evaluation of that hypothesis itself requires (*inter alia*) bringing the property *green or triangular* before the mind as such. (Fodor, 2008, p. 139)

So, for a concept-learning system to work, there need not only be the right kinds of cognitive capacities, like the ability to make inductive inferences, but also the right kinds of elements, or building blocks, to be used in this cognitive system. Whatever it is that happens in such cases of inductive inferences from hypotheses, it would be misleading to call it learning. In this way we arrive at the conclusion (9), and have reduced the common-sense notion of learning to a paradoxical process without explanatory value for a theory of cognitive development – at least in Fodor's view.

2.4.3 (3) The representational theory of mind (RTM)

Fodor directs his argument at Empiricist theories of concepts, as already made explicit above. To correctly evaluate the scope of Fodor's argument, it is important to locate it more firmly in the historical context of the Early Modern debate between Rationalists and Empiricists, and to link Fodor's notion of 'concept' to their notions of 'ideas'. Two main Empiricist theories, Locke's and Hume's, trade in 'ideas' as the main materials of thought. Hume has the additional dichotomy of impressions and ideas, both of which can be simple or complex, but for now this extra piece of theory shall not concern us. As already introduced above, Locke holds that the contents of thought arrive through perceptual means. Human perception picks out certain properties of worldly objects, and transfers them into simple ideas. These simple ideas, in turn, are the basis for complex ideas – ideas that are formed by joining at least two simple ideas together. This simple picture gives us two types of representations that we find again in Fodor's theory: simple, or atomic, concepts, and complex concepts. The distinction between simple and complex concepts should however not be confused with the distinction between lexical and phrasal concepts, as introduced above. RED, for instance, is both a simple and a lexical concept – it cannot be divided into any other constituting concepts (thus, a simple concept), and it is represented by a single word (thus, a lexical concept). On the other hand, SPARROW is a lexical concept but not a simple concept, because it is presumably constituted by a variety of defining or characterising concepts, like SMALL, PASSERINE, BIRD, or some such combination, which might even be constituted by even simpler concepts. Phrasal concepts like THE ONLY LIVING BOY IN NEW YORK are by default also complex concepts.

To elucidate the argument for the paradox further, one has to say a few more words about Fodor's representational theory of mind (RTM) since this is at the centre of his general philosophical theory, and also widely held in the cognitive sciences. A short characterisation of RTM is the following two-step definition by Fodor (1987):

> Claim 1 (the nature of propositional attitudes): For any organism O and any attitude A towards a proposition P, there is a ('computational'/ 'functional') relation R and a mental representation MP such that
> MP means that P and
> O has A iff O bears R to MP (Fodor, 1987, p. 17)
>
> Claim 2 (the nature of mental processes):
> Mental processes are causal sequences of tokenings of mental representations. (Fodor, 1987, p. 17)

Let me spell out claim 1 first. Put like this, there is a mirroring relation between some mental processes on a personal, or conscious, level, on the one hand, and some metaphysical entities that are biconditionally related to these mental processes, on the other hand. Fodor talks about an organism O, which I will for now assume to be a fellow human being, Otto. Otto has certain beliefs about states of affairs – which Fodor calls attitudes A towards some proposition P – let's say he believes that (A) grass is green (P). This is the part of the theory that describes the personal level of Otto's belief. The metaphysical picture that has to obtain just in case the statement 'Otto believes that grass is green' is true is as follows: Otto is required to have the mental representation of grass being green (MP), to which he stands in a certain relation R of accessing, having, or using it in the way that is characteristic of a belief. Different propositional attitudes are correlated to different relations R – R(believes) differs in some respect from R(fears), for example. This mental representation, in turn, 'means that' grass is green (P), which one can take to mean that, without the mental representation 'grass is green', the proposition 'grass is green' wouldn't mean what it does.

Claim 2 adds the details about how a single mental representations become parts of whole mental processes, like strings of thoughts: one can call a set of thoughts a mental process if there is a causal relation between the acts of relating to mental representations. If Otto at first thinks that (A) grass is green (P), he will simultaneously enter in the relation R(thinks) to the mental representation 'grass is green'. He might have a prior belief about cotton-grass, such as 'Cotton-grass is grass'. His first thought might now cause him to doubt (A) whether cotton-grass really is grass (P) (since cotton-grass has a white flowery top), putting him in the relation R(doubts) to the mental representation 'cotton-grass is grass', which in turn causes his wish to look up cotton-grass in his botanic encyclopaedia, etc., until he has reached an end for his enquiry.

The main thesis of Fodor's brand of representational theory of mind is also expressed in premise (3) above – thinking about something is bringing it before one's mind under a certain description, or in a certain form. This certain form is symbolic, or representational: it conveys the content of the thing thought about, yet it only *re-presents* the original thing. This re-presenting typically happens in a different medium than, say, the perceptions of the things around one, like a symbol-system that is in some way reliably linked to the things one thinks about. To apply this to the concept-learning example above, in order to think about green or square things, one needs to mentally represent these things as green or square. And this, in the HF story, is an inductive, inferential process:

1. If something is green or square, then classify it as GREEN OR SQUARE.

2. (Inductive step) If x is green or square, then all things that are green or square belong to the same category (categories) as x.

3. x is green or square.

4. Therefore: x belongs to the category (categories) GREEN OR SQUARE, and all other things that are green or square belong to the same category (categories).

Forming these thoughts requires one to represent a certain state of affairs as such a state, under the correct description. This, in turn, requires one to have the means to represent the state of affairs that way – one needs the mental vocabulary, or the access to the categories, or the nomological mind-world relation, or whatever one would like to call it. The correct description must be understandable to the learner. Also, as Fodor (1981) remarks, to form such thoughts, the thinker must have "an a priori ranking of the hypotheses in order of relative simplicity" (Fodor, 1981, p. 268) and

> a confirmation metric: roughly, a function from pairs consisting of hypotheses and data sets onto numbers that express the level of confirmation of the hypothesis relative to the data. (Fodor, 1981, p. 269)

All of the elements identified in this paragraph are required for proper induction-based learning, as in the HF model. So, for any kind of learning to happen, a lot of 'mental machinery' in the form of an "inductive logic" (Fodor, 1981, p. 269) and of a vocabulary for that logic must be available prior to even the very first episode of learning.

In Fodor (1981), Fodor had still held that at least the formation of complex concepts, like FLURG, there can be concept learning – this learning being the combination of innately given primitive concepts to give their combination a new label, or a specific use in a given task. But this kind of learning isn't of much interest to him, precisely because of the assumption that only simple concepts can be at the basis of the conceptual system. This idea stems from his assumption that simple (primitive) concepts are for the most part concepts that are triggered by the 'sensorium', i.e., the human sensory system. If one adheres to this atomist premise, it's just natural to have complex concepts playing a minor role in a theory of the origins of concepts because one would assume that all further concepts – all complex concepts – must be combinations of the innate simple concepts. Furthermore, as the *LOT2* argument attempts to show, it doesn't matter whether the concepts are simple/primitive or complex. The earlier argument in Fodor (1981) had the further empirical premise that "quotidian concepts are mostly primitive" (Fodor, 2008, p. 129f.), which amounts to saying that most of these concepts cannot be decomposed into other concepts; they are atomic in structure.[2] In Fodor's own words, he used this additional empirical premise for the following argument:

> If one then throws in the (empirical; see above) assumption that most of the concepts one has *are* primitive (which is to say, not definable) you get the consequence that most of the concepts one has can't have been learned. (Fodor, 2008, p. 137f.)

So, the conclusion back then was that primitive concepts must be innate while complex concepts could be learnt. Now, in the *LOT2* argument, all concepts are innate or not learnt, as concluded in (9).

2.4.4 (9) The anti-learning conclusion

In the face of his paradox, and with the additional reasoning laid out above, it might already be clear why Fodor draws the consequence that all concepts must be either innate or non-rationally acquired. The HF model cannot explain the learning of primitive concepts, which have in any case been assumed to be innate by both Empiricists and

[2] I will not delve into Fodor's argument for Conceptual Atomism, or into his views on the definability or structure of our quotidian concepts, and just assume his view is right for the sake of the argument.

Rationalists alike. But Fodor (2008) argues for the even stronger conclusion that complex concepts aren't learnt by hypothesis-testing-and-confirmation. Since the primitives from which a complex concept is composed aren't learnt, the only thing that happens in HF model learning of a complex concept, say GREEN OR SQUARE, is that a string of already available concepts is tested as the label for a special class of cases in which the concept might apply. Putting together old building blocks doesn't create anything previously unavailable, and therefore isn't to be called learning, or so Fodor would argue. This doesn't mean that one cannot learn new facts, or get new beliefs, about things one has already been acquainted with – the concept FLIGHTLESS BIRD might have no extension for you until you learn about penguins and their earth-and-sea-bound life. But the concepts that you need to think about flightless birds must have been in your cognitive system before you have learnt about penguins, so the acquisition of new beliefs is contingent upon the possession of the necessary concepts for forming that belief.

For many, this conclusion will certainly be undesirable, but for a large part of his career, Fodor accepted it and made it one of the foundation stones of his Radical Concept Nativism. With this move, he avoided the problems that he identified for theories of concept learning. Yet, he had to give an account on his view of conceptual development. It is one thing to say that all concepts are innate, but it is another thing to say that all concepts are known, or available, at birth. This latter claim is stronger, since it presupposes a strong sense of innateness of mental contents: by that reading, a newborn would *know* what a carburettor is. Fodor wants to avoid this strong reading and its absurd consequences; therefore, he has to tell a story about how experience with the things one has concepts of leads to the activation of those concepts as parts of the cognitive vocabulary. How does a child's experience with cars lead to her coming to know what a carburettor is?

How is the innate conceptual repertoire activated? How do primitive perceptual concepts, like colour concepts, get to designate red things in the person's experience? Fodor's thoughts on this topic have evolved since Fodor (1975), and I will trace this development in the following and then in section 2.5 of this chapter. While Fodor's earlier writing on the topic has been characterised by Radical Concept Nativism, from Fodor (1998) on, Fodor abandons Radical Concept Nativism, and advocates a more moderate Nativist position that doesn't rule out that some concepts are acquired, although still unlearnt.

In the first step of the development of Fodor's argument, the notion of 'triggering', as developed in detail in Fodor (1981), is introduced as solving the problem. There, he claims triggering to be the only way to acquire a concept. Strictly speaking, the concept is already there, in the wealth of symbols in the Language of Thought (Fodor, 1975), but it hasn't been activated yet – the person may have the concept, but it's not implemented in the actually used mental vocabulary. Triggering makes the connection between a word, or an experience, or anything else from outside the cognitive system, and the concept – the concept gets triggered, or activated, by the occurrence of said word or experience. The important thing is that triggering is not a rational process and can, in one sense, happen arbitrarily on any stimulus – it is a "brute-causal" (Fodor, 1981, p. 280) mechanism, like switching on a lamp in a previously dark room. Yet, in order for the triggering story to succeed, Fodor needs a story that reliably links a trigger to a concept without having the triggering experience carrying conceptual content/being under the control of a rational process. This is the point of a non-rational account of concept acquisition: how do brute-causal events influence the course of our mental development – which is also the development of our rational capacities?

Consider an example for a triggering experience: newborn ducklings imprint on their mother, which means they get attached to their mothers. The dominant cause of attachment is the movement of the mother. This makes it possible to exchange the mother for another animal, or even an automatically moving object, and still observe the imprinting effect – the duckling will follow that other animal or object, as long as it is the nearest moving object the ducklings perceive.

As Fodor claims regarding this example, given that movement nearby is a quite reliable indicator for a newborn for having the mother nearby (and so, there's an evolutionary explanation for this kind of imprinting mechanism), triggering is not random. But then, there could have been other evolutionary solutions to achieve imprinting just as reliably, like 'looking like a duck', or emitting a certain smell; yet, movement is the 'evolutionary chosen' one. So, one might say there's some randomness in actual triggers within the boundaries of an evolutionarily given set of alternative triggers. Until Fodor (1998), he was content with this basic account of concept acquisition – if it isn't innate, there is some kind of innate mechanism that will come online once the right kind of object shows up. As I will explain in the next section, Fodor has further developed this rough picture after he identified the so-called doorknob/DOORKNOB problem. As a reaction to this problem, he has developed two theories that go beyond the simple triggering account I just presented while retaining the non-rational aspects of triggering.

2.5 The doorknob/DOORKNOB problem

While Fodor has no qualms with his argument's result that concepts cannot be learnt, he recognises that his position faces a problem when he wants to explain the relation between experiences of given objects and the acquisition of the concepts pertaining to those objects. His example in Fodor (1998) is the relation between experiences of doorknobs and the concept 'doorknob'. Because of this choice of example, he calls the problem the 'doorknob/DOORKNOB (d/D) problem'. In Fodor's own words, the d/D problem takes the following form:

> Why is it so often experiences of doorknobs, and so rarely experience with whipped cream or giraffes, that leads one to lock to *doorknobhood*? (Fodor, 1998, p. 127)

The HF model is a solution to this problem, since it is based on an evidential relationship between concepts and their referents. But the HF model cannot work because of Fodor's paradox. The brute-causal acquisition story has to find an alternative solution to this problem: how can it guarantee that a concept is triggered by the right kind of object – the object that the concept is supposed to refer to?

2.5.1 The metaphysical solution

In Fodor (1998), Fodor discusses two metaphysical theories that could potentially solve the d/D problem. The first is to adopt a causal/historical theory of reference (cf. Fodor, 1998, p. 133). There are several kinds of causal/historical theories of reference to choose from, all of which could explain the correctness of acquiring DOORKNOB from experiences with doorknobs by pointing out the causal or historical role doorknobs play in the process. Fodor has his own version of a causal theory of reference, asymmetrical dependence theory (cf. Fodor, 1990), to offer. However, he rejects the solution of appealing to causal/historical factors because he thinks this requires a mechanism like the HF model to work (cf. Fodor, 1998, p. 133), which would defeat the purpose of providing a non-learning solution to the d/D problem. He explains this in the following way: a causal/historical theory of reference would have to appeal to more than just causal interaction with a doorknob to get DOORKNOB, and this additional bit of needed explanation can only be a psychological mechanism which ensures the acquisition of the right concept. Since the only such mechanism Fodor acknowledges is the HF model, causal/historical theories of reference cannot solve the d/D problem, as their solution leads them back into Fodor's paradox.

Instead, Fodor proposes a second kind of metaphysical solution to the d/D problem that hinges on the mind-dependency of those kinds of conceptual content that one would assume to 'get from evidential experiences' with the concepts' referents. The mind-dependency Fodor envisages here is analogical to the mind-dependency of colour and other appearance concepts. According to this idea, objects aren't coloured themselves, but the colour we perceive is dependent on our specific perceptual make-up. For instance, if you are red-green colour-blind, you will perceive tomatoes and cucumbers as having roughly the same colour, whereas people without this condition will see clear colour differences between the two. This isn't because the objects change dramatically depending on who looks at them, but because of differences in the onlookers' perceptual representations. So, if concepts like DOORKNOB are like appearance concepts in this sense, then some specific features of the human mind make doorknobs appear *like doorknobs* to us. Fodor puts it this way:

> *being a doorknob* is having that property that minds like ours come to resonate to in consequence of relevant experience with stereotypical doorknobs. (Fodor, 1998, p. 137)

This solution works for Fodor because it relies on the triggering idea and solves the doorknob/DOORKNOB problem: you get DOORKNOB from experience with doorknobs, and not from experience with giraffes, because 'minds like ours' are put together in such a way that experiences with doorknobs trigger that concept.

Jesse Prinz (2002, 2011) criticises Fodor's solution for the d/D problem, and gives several reasons for not accepting his argument that 'doorknob' is a mind-dependent concept in the way Fodor intends it to be. The first reason is that there are crucial differences between concepts that are typically regarded as mind-dependent, such as colour concepts, and concepts like DOORKNOB:

> Properties like red are presumed to be mind-dependent because it is difficult to find anything that red things have in common other than the kinds of experiences they cause us to have. (Prinz, 2002, p. 232)

As Prinz continues to point out, for things that don't fall under appearance concepts, it is usually very easy to find things that they have in common, such as a common function – if it is used to open doors, it *ceteris paribus* might well be a doorknob. We don't decide to conceptualise something as a doorknob because of the way it looks to us, but because we see good reasons to treat it as a doorknob. Without any further reasons to liken doorknobs to instances of the colour red, the analogy turns out to be quite weak. Consequently, Prinz's verdict is that Fodor's notion of 'mind-dependency' is unwarranted, and should therefore be rejected.

The second reason Prinz gives for rejecting the metaphysical solution is that it excludes at least some instances of a concept from triggering it, while including some things that are not actually instances of the concept (cf. Prinz, 2002, p. 233). In the first class of things, you would find atypical doorknobs that wouldn't trigger DOORKNOB. Imagine, for instance, a doorknob that is shaped like a lawn gnome. According to Fodor's view, it shouldn't activate the concept DOORKNOB (it shouldn't *strike us* to be a doorknob), but would it activate LAWN GNOME? In the second class of things, we would find images of (typical) doorknobs, or little figurines of doors with immovable doorknobs, and similar objects. They are all representations of doorknobs, but aren't the right kind of object to trigger DOORKNOB. That role is reserved for actual (typical) doorknobs.

In sum, the metaphysical account of regarding doorknobhood as a mind-dependent property that only gets activated by typical instances of DOORKNOB faces serious difficulties, against which Fodor doesn't adequately shield it.

2.5.2 The neurological solution

In later works (Fodor, 2008), Fodor has gone on to develop a biological, or neurological account of concept acquisition at which he had only hinted when he introduced the metaphysical account in Fodor (1998). This still incorporates the main idea of the metaphysical solution – that concepts have the contents they do because their referents make our minds trigger the right kinds of concepts – but adds an acquisition story to the plain metaphysical picture. Margolis and Laurence (2011) describe the upshot of this solution in the following way:

> Fodor is making the far stronger claim that concept acquisition is subject *only* to neurological explanation. If Fodor is right, cognitive psychology is no more relevant to the study of concept acquisition than it is to the study of red blood cell production or digestion. (Margolis and Laurence, 2011, p. 531)

According to the neurological solution to the d/D problem, concept acquisition is a two-step process. In a first step, which Fodor calls 'P1', a stereotype pertaining to doorknobs is learnt from experience with doorknobs. I will represent stereotypes in bold typeface – stereotypical doorknobs would give a learner the stereotype **doorknob** on Fodor's view. 'Stereotype' is Fodor's word for what the psychological literature refers to as 'prototype', namely a complex mental representation that is built with the help of statistical knowledge about instances of the category. I will use 'stereotype' in the context of Fodor's theory and 'prototype' when talking about the psychological theories. Fodor specifically posits that stereotypes are learned by an inductive process, like the one at the heart of the HF model (cf. Fodor, 2008, p. 153). That means that the learner learns a stereotype of a given object of her experience, for instance of a sparrow, by forming a statistical representation **sparrow** through observation of typical sparrows. The process is thus evidence-based and inductive, as the prototype will be expected to apply in future encounters with sparrows. All aspects of concept learning that Fodor finds problematic in the context of learning *concepts* – the evidential relation to instances and the reliance on inductive inference – are thus covered by stereotype learning. This is supposed to relieve concept learning of the problems Fodor's paradox creates, while still giving newly acquired concepts grounding in experience. In *LOT2*, Fodor doesn't consider the possibility that his anti-learning argument might also apply to the learning of stereotypes; this is a point I will return to when raising critical points against Fodor's neurological model of concept acquisition.

In the second step, which Fodor calls 'P2', a neurological process that is not under conscious or intentional control takes over, and creates a concept from the stereotype. Fodor claims that this neurological process cannot be construed as an inference, and that it therefore is not an inference-based account of learning. All that needs to be present to gain a concept of an experienced object is a stereotype that was learnt from these experiences, and a brute triggering of the neurological concept-formation process. He admits that there are empirical gaps in his story, but proposes them as interesting directions for future research (cf. Fodor, 2008, p. 168).

This solution avoids the d/D problem because it grounds the concept in experience, by the mediating help of the stereotype: the stereotype **doorknob** is learnt from experience, while the concept DOORKNOB is still not learnt. Fodor summarises his point in the following way:

> What's learned (not just acquired) are stereotypes (statistical representations of experience). What's innate is the disposition to grasp such and such a concept (i.e. to lock to such and such a property) in consequence of having learned such and such a stereotype. (Fodor, 2008, p. 162)

One strength of Fodor's neurological solution is its acceptance and use of the psychological data on prototypes in categorisation. He has continuously argued that prototypes cannot be concepts because prototypes don't fulfill several of his core demands on what concepts should do; for instance, prototypes cannot combine in the way concepts can – they are not compositional (cf. Fodor, 1998), but now he has found a place for them in his theory of concepts after all. Also, the neurological account seems to fare better against some of the criticism that Prinz raised against the metaphysical account. For instance, the first step of the acquisition process can provide explanations of how non-typical instances, or non-instances of a stereotype can still be conducive to learning the right kind of stereotype. For example, pictures of typical doorknobs can bring about the right kinds of inductive inferences for learning **doorknob** because of their similarity to the real object.

However, Margolis and Laurence (2011) give strong arguments against accepting Fodor's solution. The first point of criticism Margolis and Laurence raise is that the two-part story of stereotype-learning plus neurological concept wiring faces the same paradox as standard 'cognitive psychological' concept learning. The problem stems from Fodor's insistence that learning has to be based on the formation and testing of hypotheses. If we transpose the HF model for learning to stereotypes, then we still face Fodor's paradox. This is the first horn of the dilemma Margolis and Laurence (2011) raise. The second horn immediately comes into view in the case of the Fodorian offering a different kind of learning mechanism for stereotypes: if it can be used for stereotypes, then it is likely that it can also be used for concepts. Thus, Fodor would defeat his own argument if he allowed for a mechanism of stereotype learning that differs from the HF model.

At this point, a critical reader might already raise the following point: if we can use Fodor's paradox against his own argument for a mechanism of concept acquisition, then why shouldn't we just stop considering Fodor's position, and his argument against concept learning itself? This is actually foreshadowing the solution that I will develop in Chapter 7. I will postpone discussion of this point until then in order to build up a complete case against Fodor. Furthermore, we shouldn't dismiss Fodor's paradox just because the solution he offers is not satisfactory – the problem itself might be deeper than he expected. For example, Fermat's Last Theorem wasn't proven for centuries, and is considered very important in mathematics even though Fermat himself didn't offer a (successful) proof.

The second critical point for Margolis and Laurence is that concept acquisition actually needs a psychological component, contra Fodor's explicitly stated programme. They give three reasons for believing so:

1. Concept acquisition is strongly influenced by personal preferences and cultural factors. Across different cultures, the same stereotype can trigger many different concepts, as evidenced by differences in colour categorisation across cultures.

2. Many concepts are acquired without a stereotype because there are no stereotypes for the things they represent. Prime examples are complex concepts, like SIMON AND GARFUNKEL SONG or HEDGE FUND, and abstract/unobservable concepts, like ELECTRON or DEMOCRACY.

3. Some concepts are learnt by psychological mechanisms other than stereotype formation. Good examples for this include learning concepts by using a theory (consider the case of GRAVITY), or by applying psychological biases such as the essentialist bias.[3]

Margolis and Laurence argue that Fodor cannot account for these factors without letting psychological explanations slip back in, thus defeating the purpose of the move down to the neurological level. On the grounds of these serious objections, I conclude that Fodor's explanations of concept acquisition without the influence of psychological states fails.

2.5.3 The upshot of Fodor's position

This is the outlook we are facing after considering Jerry Fodor's position on concept acquisition: concepts cannot be learnt because the HF model leads to a paradoxical situation. But Fodor's proposals for brute-causal concept acquisition, whether metaphysically or neurologically motivated, also fall flat in the face of strong criticism. If we want to follow Fodor and agree that his Learning Paradox is a serious problem for theories of concept learning, then we are in deep trouble. Should we conclude that, based on our investigation so far, not even brute-causal acquisition is possible? Or should we look for other ways of saving brute-causal acquisition? I want to propose that we should a) not focus on the brute-causal path and that we should b) still take Fodor's paradox seriously. Regarding (a), we should take another careful look at concept learning: if we found a way of demonstrating that concept learning is possible after all, then we might not even have to come up with other types of acquisition mechanisms. In the next section, I will show how the search for concept-learning mechanisms can be put in motion from the background of Fodor's engagement with it. Regarding (b), I want to note two points: first, Fodor's paradox is only the most recent example of a philosophical position with a long intellectual history. Versions of it have been held throughout the history of Western philosophy. We shouldn't discard this strong Platonist/Rationalist tradition without a thorough discussion of its merits and its weaknesses. Second, philosophers have only fairly recently come into a position which allows them to engage with the empirical aspects of the study of the mind. The cognitive sciences have brought a wide range of new methods and results to bear on philosophical questions. In order to do justice to philosophical positions such as Fodor's, it is essential to actually confront them with a range of scientific evidence before we reach a verdict on their soundness.

[3] I will say more about these two options in Chapters 3 and 6, respectively.

2.6 Fodor's Challenge

Fodor's paradox itself is an argument against the possibility of concept learning. In the remainder of the thesis, I will investigate this paradox. The argument poses serious difficulties for any theorist who wants to contest it. Anybody who wants to maintain a notion of concept learning or who wants to offer a concept-learning mechanism (CLM) that doesn't succumb to Fodor's paradox, will have to answer what I will call 'Fodor's Challenge':

Fodor's Challenge To overcome Fodor's paradox, one has to propose a concept-learning mechanism that cannot be reinterpreted as a version of the HF model without significant explanatory or theoretical losses.

Fodor repeatedly opines that his take on what concept learning is is the only game in town (cf. Fodor, 1975, 2008). For instance, he accuses those defenders of concept learning who claim to offer an alternative to the HF model of just not realising that their model actually is an instance of the HF model (cf. Fodor, 2008, p. 132). Now, if all other accounts fail because they can't resolve the learning paradox (because they are just instances of the HF model), one has to accept Fodor's take on the issue. This way of arguing is an instance of an inference to the best explanation (IBE) – when comparing all the alternative explanations, the HF model purportedly comes out ahead of the challengers.

The important point regarding an argument from an IBE is that such an argument can always be challenged by providing an equally coherent alternative theory: if there are two theories, both of which giving an *equally satisfying* reply to the paradox, the race is open again. At this point, the virtues of the two theories have to be weighed against each other to guarantee a fair and qualified inference to the best explanation. As the literature on IBE and on scientific explanations offers a variety of standards of explanatory quality (e.g. Lipton, 2004; Van Fraassen, 1977; Thagard, 1978), I want to only briefly mention two classes of relevant virtues: empirical virtues and theoretical virtues. First, a theory of concept learning can be better than others because it fits better with available empirical data, or because it allows for better predictions concerning future experiments. Second, a theory of concept learning can be better than others because of its "scope, precision, mechanism, unification (...) [or] simplicity" (Lipton, 2000, p. 187).

As a final point, if one of the competing theories is able to save the notion of learning from being obsolete, as in Fodor's theory, that theory would have one important benefit – it would accord to everyday human experience. When watching children growing up, we see them learning new things all the time; we see them making mistakes, and then getting things right; we teach them, and they explore the world on their own as well. To put it bluntly, the question for the layperson is not whether children learn concepts, but how they do it. In the following, I will take these considerations as the measuring pole for how well alternative theories of concept learning do when facing Fodor's challenge.

Now that I have identified an important problem for theories of concept learning, and the challenge it imposes on the field, I can begin the investigation of the contenders for answering the challenge. The first set of concept-learning mechanisms I will consider have been proposed by developmental and cognitive psychologists.

Chapter 3

Developmental accounts of concept learning

3.1 The challenge for developmental psychology

In the previous chapter, I have introduced Fodor's challenge for theories of concept learning, and have argued that it poses a universal threat to the very idea of their success. I will now go ahead and look at proposals to overcome this challenge, starting with developmental psychology in this chapter and moving to cognitive psychology more generally in the next.

For developmental psychologists, Fodor's challenge has special importance. After all, if concept learning (CL) were possible, it would surely happen a lot in early ontogeny, and should be studied extensively. There are several models specially directed at explaining CL in early cognitive development, and several important books have been devoted to conceptual development, conceptual change, and concept learning (Carey, 1985, 2009; Keil, 1989; Gopnik and Meltzoff, 1998; Mandler, 2004). Among this group, Carey and Mandler stand out because they have explicitly addressed Fodor's work and proposed ways to avoid his anti-learning stance. In this chapter, I want to raise a problem for this kind of research. The starting point and guiding question are as follows: especially in prelinguistic infancy, there are only very few experimental paradigms that afford data pertinent to infant cognition. How strongly can this data support a theory of concept learning?

In the following, I will give a brief review of the current state of research in infant concepts and of the methods of acquiring experimental evidence. Based on this, I will argue that the evidence fails to support the available theories of prelinguistic concept learning. Then, I will discuss Jean Mandler's and Susan Carey's models for concept learning and will investigate whether this argument affects them. My conclusion will be that Mandler cannot escape it given the present state of infancy research, while Carey is in a better situation but might not be able to explain prelinguistic concept learning through her model.

3.2 The 'Evidence for Prelinguistic Concept Learning' hypothesis

The vast improvements in methodology and in technology in the past decades have brought about a great amount of research in infant cognition, which propelled developmental psychology far away from the modest empirical data the likes of Vygotsky (1962) or Piaget (1952, 1954) could generate. Contemporary research still has one main limitation in the fact that the study of infant cognition is constrained by the performative abilities of its young experimental subjects.

Still, psychologists have found ways to circumvent these limitations and have developed ingenious methods to better understand how young children think. In the infant concept learning literature, researchers typically use the following three experimental paradigms, which I will introduce before developing my argument against the claims for the evidence's relevance for concept learning.

3.2.1 The dominant research paradigms in early infancy

Looking time studies

The most important paradigm in infant cognition research arguably is the looking time, or dishabituation, paradigm, which takes infants' looking times as evidence of their interest, or surprise, when confronted with a novel stimulus. Looking time studies have an important advantage to the other methods I will introduce below: since they rely only on infants' ability to direct their visual attention to a given display, they can be done with very young participants, starting around three months of age (as for instance in Quinn et al., 1993, see below).

In studies using this paradigm, infants are presented with a visual display, and the times they take to look at aspects of the scene, e.g. different parts or different objects, are measured. The researchers continue to measure infants' looking times during familiarisation (habituation) periods until the test trial, in which the infant is supposed to dishabituate. Dishabituating is the act of an infant looking considerably longer at a visual scene in front of her in cases in which she recognises a novelty or a difference to previous trials.

Murphy (2002, p. 274) describes the looking time paradigm as the main method of infancy research, and Mandler (2004) concurs.[1] It has the virtue of bringing novelty effects, or surprise, to the fore. By this, a dishabituating child can 'tell' the researchers about her expectations with regard to what she expected to happen next. If the test stimulus appears to be similar to the habituation stimuli in relevant respects, then the child will continue to habituate, i.e., to look for a shorter period of time. However, if the infant finds a new stimulus sufficiently novel, surprising, or different in a relevant respect to previous displays, then this will be reflected by longer looking times.

Quinn et al. (1993), in one important contribution to the infant categorisation literature, use the looking time paradigm to test the acquisition of new animal concepts. Infants aged 3 to 4 months are presented with photographs of pairs of animals, mainly pairs of dogs or pairs of cats, in six habituation trials. In the test trial, they are confronted with a picture of a dog or a cat, and a picture of a bird. The bird, as the novel stimulus, is looked at considerably longer (combined mean preference 62.64%, see Quinn et al., 1993, p. 468) than the by now familiar cat, or dog. Quinn, Eimas and Rosenkrantz take this as evidence that the category *cat-or-dog* has been formed during the habituation period:

> The preference for novel exemplars from a novel category is consistent with the view that during the familiarization period the numerous instances of dogs, for example, came to be represented by a single categorical description. (Quinn et al., 1993, p. 468)

[1] The paradigm is used successfully in several other areas of infancy research besides categorisation behaviour, for instance by Renée Baillargeon and colleagues (Baillargeon and DeVos, 1991; Onishi and Baillargeon, 2005).

In subsequent experiments, also reported in Quinn et al. (1993), the researchers aimed to exclude the possibility that infants prefer seeing birds to seeing dogs or cats, and the possibility that they weren't able to discriminate between the category members of the cats/dogs category.

Touching and examination

A second major type of experimental evidence in early childhood categorisation comes from studies that measure more than just looking times – namely, groping and touching of objects. This method is only available for slightly older children, given that very young infants have limited control of their limbs (Mandler, 2004, p. 10). In one version of the task, the sequential-touching paradigm, researchers present infants with a set of toys, for instance animal toys, and go on to record the order in which the infants touch the toys. This is based on the observation that the children 'group' or 'categorise' the objects by touching them in sequence – they would first touch all dogs, then all birds, then all elephants, etc. (Mandler, 2004, p. 9).

Another instance of the paradigm, the object-examination task, is similar to Quinn's version of the dishabituation paradigm, in that it presents infants with pairs of like toys, and measures the length of the tactile engagement with the toys, up until the test trial, which would feature one member of the 'old' category and one member of a 'new' category; it could be dogs and birds, as above. If the infants engage much longer with the new kind of toy, this could be seen as evidence that they regard the new toy as of a new category, and that they have learnt the category represented by the old toy. They learnt DOG from touching the toy dogs, and are surprised by the toy bird, which they don't group under the DOG concept (Mandler, 2004, p. 9).

Imitation

A third kind of paradigm makes use of infants' propensity to imitate actions that they witness around them. It helps illuminate several aspects of infant cognition, such as categorisation choices and inferential reasoning. Just like the dishabituation paradigm, it is used in various subfields of infancy research, very notably in Gergely and Csibra's social cognition research (e.g. Gergely and Csibra, 2003, 2005).

Jean Mandler uses two kinds of imitation paradigms, generalised imitation and deferred imitation. In generalised imitation (Mandler and McDonough, 1996), 1-year-olds are first shown certain actions, and then given the opportunity to imitate them, with slight changes to the setting. One of Mandler's examples is giving a drink to a toy: children first observe the experimenter as she gives a dog a drink three times in a row, and then are offered the cup and a choice of two toys, like a car and a bird. Based on their choice of imitative activity, researchers can draw conclusions about whether the experiment's participants have generalised from the fact that dogs drink to the fact that animals drink. With this kind of experiment, Mandler gathers information about the categorial hierarchies that infants have established at the age of 1. Furthermore, she regards it as a test of the limits of perceptual similarity (Mandler and McDonough, 1996, p. 315f.), since sometimes perceptual similarity would override a judgment made along category lines – if children treat penguins and eagles alike in a task, but treat eagles with extended wings differently than airplanes in the same task, this can be regarded as evidence for inferences based on categorisation rather than similarity.

In deferred imitation, infants are presented with a sequence of events, and after varying delay lengths, are allowed to imitate the event they witnessed. In one study (Mandler and McDonough, 1995), 11-month-olds were shown causally related and unrelated events, and allowed to imitate after 20 seconds, 24 hours, and 3 months. Examples of imitated events include putting together two pieces to form a rocking horse and rocking it (causally related), and feeding a toy rabbit a carrot and giving it a hat (causally unrelated). In this study, infants performed much better when imitating the causally related events after 3 months, despite being able to do both kinds of imitation well after 24 hours. The deferred imitation paradigm gives insight into the kinds of features of an event or an activity that the child remembers after the delay. This is important because it can be regarded as evidence for the exercise of conceptual abilities. Mandler holds a strong position on this, claiming that "[y]ou can't recall anything you haven't conceptualized" (Mandler, 2004, p. 10). If one accepts this position on the relation between memory and concepts, then deferred imitation studies indeed provide a way of testing conceptual capacities in infants.[2]

The EPCL hypothesis

With all these virtues of the experimental paradigms in early ontogeny research being accepted, it seems reasonable to assume that this kind of research could also tell us about conceptual development. I propose to call this the 'Evidence for Prelinguistic Concept Learning' (EPCL) hypothesis:

EPCL Experimental results from studies with prelinguistic infants can provide evidence for concept learning.

As my main example of adherence to this hypothesis, I will use the work of Quinn, Eimas and Rosenkrantz. Quinn et al. (1993) take their evidence from dishabituation behaviours to support the hypothesis that concepts are learnt in the process of participating in a looking time study, such as described above. They do not specify a learning model, but refer to perceptual differentiation (cf. Quinn et al., 1993, p. 473f.) as playing a part in the acquisition of novel categories.[3]

[2] We will shortly return to the issue of what counts as conceptual, especially in relation to perceptual representations. For now, I want to offer Mandler's view as a justification for thinking that this research paradigm has potential for research in infant concepts.

[3] In the article in question, the distinction between 'perceptual categories' and 'conceptual categories' is not entirely straightforward, and it isn't clear whether the acquisition of a category involves the acquisition of a concept. Gauker (2005) makes a similar point about the work of Quinn and colleagues.

In the following, I will present an argument against the EPCL hypothesis generally, assuming that Quinn's work is an adequate illustrating example.

3.2.2 The argument against EPCL

I take the following to be quite uncontroversial assumptions about the demands on a mechanism for concept learning, which lead to a conclusion current theories of infant concept learning have difficulties escaping. I will go through the individual assumptions and the evidential basis for them right after the following outline of the argument.

1. Concept learning is a process in which the crucial steps of a learning event happen between an initial moment that starts the process, and an end point, at which a concept has been learnt.

2. If a set of empirical results only captures start and end states, then – barring strong abductive pull towards one explanatory option, and all else being equal – the explanation of the process is only weakly constrained by the individual piece of evidence. That means that there are often several equally valid proposals for explaining the process happening between start and end of the learning episode.

3. If a set of empirical results includes intermediate results, then these should be regarded as among the strongest kinds of evidence for or against an explanation.

4. The only in-between data in infant concepts studies are looking times, touching times, or touch sequences.

5. This kind of evidence underdetermines the choice of learning mechanism - it is compatible with too many mechanisms.

6. Conclusion: Barring additional evidence, infant concept learning studies don't support any concept learning mechanism.

Spelling out the argument

I regard (1) as a quite innocuous statement of fact. If there is any concept learning, then it surely starts with a point in time, t_1 at which the concept that is to be learnt is not available yet, and at which something sets a cognitive process in motion, which is related to the learning of this concept. Equally, at some future point in time, t_2, the learning process will have been completed, or will be complete enough to grant that a concept has been learnt. Whatever happens between these two points in time will be of interest to a theorist looking for a mechanism of concept learning.

This ties in directly with (2): if our investigation into a learning process doesn't yield anything besides empirical descriptions or empirical data related to the start of the process and the end of it, then we have a potentially wide range of choices with regard to explaining the process. This is especially important in cases where we only have individual results that don't (yet) tie in with, or support, other kinds of experimental data. Our test subject might not recognise flurgs as tulas at time t_1, but may do so at t_2. Now, we might come equipped with a theory of cognitive development which forces us to exclude a certain type of explanation of the learning process – for example, we might not want to explain it in terms of dynamical coupling, but insist on connectionist reweighing of certain nodes. Such prior theoretic commitments can limit the range of available explanatory hypotheses even when the empirical evidence doesn't help in singling out a process. It is however important to note that this doesn't yet speak for or against a particular mechanism: the theory behind the interpretation of the experimental data might be flawed on other grounds. Also, the data itself isn't very strongly supporting if it only covers t_1 and t_2 – if it doesn't cover the steps that lead to learning success at t_2.

Suppose now that we have found a method of getting experimental data from intermediary stages of the learning process. Suppose that we regularly ask our test subjects about their beliefs regarding tulas and flurgs, or that we measure their brain activity while engaged in tula-related tasks. These kinds of data can then be compared to the prior theoretical commitments mentioned in the previous paragraph, and can lead to a hypothesis about the learning mechanism at work. I take it that something akin to this is the best we can do *to track what goes on during learning*, as expressed in (3). It might not be the best kind of evidence for or against a given learning mechanism *tout court*: suppose that we have several kinds of converging evidence from a variety of experimental methods that all point in the direction of a given concept-learning mechanism (CLM), but which don't offer any insight into the intermediate steps happening in the process. If we now find another kind of evidence that gives us this kind of insight, but contradicts the large cluster of previous evidence, then the property of going into a finer grain of the process doesn't necessarily outweigh the clustered evidence. In such a case, other factors would have to enter into the evaluation. Still, I want to propose that researching the nature of temporal processes, such as learning, benefits from efforts to get to the finest possible temporal grain. Now, we can go on to apply the insights from (2) and (3) to the study of infant cognition. At their weakest interpretation, (2) and (3) are just pleas for temporally fine-grained experimental data in addition to the more universal demand that our data should give broad support – support based on several distinct kinds of methods, paradigms, or levels of investigation – to a theory. If we now agree that there are only the three kinds of early ontogenetic evidence that I have introduced above, then (4) is already established.

This leads to the main problem that I want to raise: just with looking times, or touch sequences, or touching times, we cannot determine what kind of cognitive process is at work in the learning episode. These kinds of evidence don't determine the associated mental processes, and don't go into fine enough detail to help excluding CLMs like the HF model. I will focus on the case of looking times now, and discuss the other two kinds of evidence in Section 3.3, as part of the critique of Mandler's theory.

The problem with studies like Quinn et al. (1993) is that dishabituation in the face of a pair consisting of a cat and a bird is just as consistent with an abstraction-from-perception story as with a hypothesis-formation-and-testing story. The infant could abstract away the defining features of cats during the habituation period, until she sees the non-catty bird and the sense of novelty of the situation leads to renewed interest after all the work of stabilising the cat concept. But the infant could just as well test her hypothesis about cats (for instance, ‘These things come in pairs’) and get bored because the result is always the same – two cats – until the cat and the bird come up, giving a different set of data to test the hypothesis on. Of course, the list of potential learning models needn’t end here – there may be many more, and I suppose that neither of them will be supported more strongly by the evidence than the others.

The only conclusion that I can propose to draw from this is to accept that the means we have to study infant cognition don’t tell us enough about the learning processes during an episode of concept learning. This does not exclude the possibility of concept acquisition, since we have experimental data that at least shows changes in the ways infants react to the objects psychologists want them to form concepts of. Infant reactions at the time t_1 differ in interesting respects from infant reactions at the time t_2. So, the question ‘Do prelinguistic infants acquire concepts?’ might be answered with ‘Yes’ even on Fodor’s watch. But my conclusion excludes the possibility of claiming prelinguistic concept learning, at least as long as the possibility of acquiring the concepts through the HF model hasn’t been ruled out by a decisive feature of the learning episode. If, ceteris paribus, an explanation appealing to the HF model is equally well supported by the evidence at hand as the alternative explanation, then there will be good reasons to prefer the HF-model-type explanation. This is so because the HF model has many theoretical virtues that make it highly attractive, such as simplicity, elegance, and parsimony of cognitive mechanisms (although not of innate conceptual primitives).

Still, importantly, this argument does not even rely on the possibility of finding another type of justification for a theory choice. It just aims to show that research in prelinguistic ontogeny doesn’t provide the right kind of evidence to determine a theory of concept learning that uniquely fits it.

Furthermore, the point of this argument is not about recognising bunnies/cats/etc., because that is possible without a learning process. What I object to on the grounds of this argument is that the dishabituation evidence is

a) evidence for one specific kind of concept learning mechanism over another,

and that it

b) is evidence for any kind of concept learning, for that matter: if it allows for a HF-model interpretation, then it might as well not be concept learning, *pace* the force of the inference to the best explanation established in Chapter 2.

Category understanding

Now, one might object that looking-time studies do in fact contribute to the understanding of concept learning: they demonstrate infants' understanding of specific categories in virtue of their intricate experimental setup. The objection to my argument thus is that I fail to give credit to the infants and to the special characteristics of the experimental paradigm. Contrary to what I argue, there is evidence of concept learning to be found in such studies after all.

In this vein, one might point out that the infants see photos of a variety of dogs or cats, and their habituating to them must be evidence that they categorise them as members of the same category. One cannot assume that they are acquainted with many different kinds of cats and dogs at the early age of just 3 months. Consider how unlikely it would be for a young infant to have already seen rottweilers, poodles, corgis, and German shepherds in the kind of detail that the experimental study affords them. By this line of reasoning, the infants can quite reasonably be taken to be grouping animals they have never seen before into one new, unifying category. However, this objection misses the point of the argument. Two things are worth mentioning in relation to this.

First, it might well be the case that the studies demonstrate genuine concept learning. The exposure to a given number of pictures of dogs, or dog figurines, or other experimental stimuli representing dogs, might bring about a rational-causal process at the end of which a great number of infants are able to group dogs (or very dog-like animals) in the new category. But just because of this, we cannot yet infer what kind of learning mechanism is at play during the episode. The data itself can only help to exclude theories that are ostensibly inconsistent with the changes in behaviour that the data represents. This will not be enough to exclude all alternative mechanisms that compete with the favourite option of the researchers doing the experiment. So, without further evidence that helps to exclude such options, we can only speculate about how a given concept was learnt in the task, if it was learnt at all.

Second, from my argumentative point of view, I can grant that it is possible for infants to group things they have never seen before into categories without any kind of concept learning taking place in the process. It is conceivable that every infant who ever successfully participated in a dishabituation study had at least *some* concepts, prior to the test, which were of relevance in categorising the trial objects. For all that observers know, the infants could group THINGS WITH BIG NOSES and differentiate them from THINGS WITH STUBBY NOSES and THINGS WITH SHARP NOSES, assuming that all the dogs in the pictures had bigger noses than cats, and that only birds had sharp noses, i.e., beaks. This interpretation doesn't even have to include the possibility of innate dog species concepts. Even without the NOSES case, the concepts at play could be as much as 'global' concepts like LAND ANIMAL and AIR ANIMAL (in the case of dogs and cats vs. birds), or something similar.

The underdetermination argument

The argument that I constructed here can be seen as a take on a classic problem in the philosophy of science – the underdetermination of scientific theories by data, or more specifically what Stanford calls the 'contrastive underdetermination' case (Stanford, 2009, §3). One could use this fact as an actual objection to my argument: if my argument were actually a clear-cut instance of the underdetermination argument, then it would just be another blow to the idea that the cognitive sciences can produce theories that are able to go beyond the (largely) behavioural data.

I think that there is one strong reason why my argument cannot be reduced to a standard underdetermination argument: the standard argument doesn't take the issue of temporal fine grain as important as the argument against EPCL does. So, while I won't deny a kinship between the problems, I want to show that my argument focuses on unique aspects of research in prelinguistic concept learning. To establish this, I will first lay out a common form of the underdetermination argument, and then present the two most important replies to it. I will then go on to establish how this differs from my own argument.

Feest (2003), discussing Cummins (1983), phrases the problem as follows:

> Psychology explains behavioral dispositions (input-output functions) by analyzing these capacities into component parts, which are themselves dispositions (i.e., little input-output functions). Any input-output function can be calculated by many different algorithms (and executed by a different sequence of intervening variables). Therefore, the correct intervening mechanism is in principle empirically underdetermined. (Feest, 2003, p. 941)

Seeing how the discussion revolves around issues in the cognitive sciences, I will take this as one of the most general forms of an underdetermination argument relevant to the current discussion.[4] If we build our theories of human cognition on behavioural data, then we will have a wide array of theoretically possible mechanisms that happen between an 'input' (like a perceivable scene) and an 'output' (like a behavioural reaction to such a scene). Bechtel and Abrahamsen (2010) phrase the problematic aspect of this as follows:

> Empirical evidence can rule out specific models, but usually cannot decide between competing architectures. (Bechtel and Abrahamsen, 2010, p. 332)

The typical replies that are brought up to save the cognitive sciences from the grave consequences of the general, contrastive underdetermination problem are as follows. First, one can offer the greater pragmatic virtues of one theory over another as a reason to use the first one. Among these virtues, simplicity, generality, explanatory fit, and elegance are among the most important. These are brought up in many discussions

[4] Influentially, Van Fraassen (1980) has also made an argument regarding the underdetermination of theory by data, suggesting as its implication that philosophers of science should stop looking for true theories, and focus on the *empirically adequate* ones. I will however only discuss underdetermination as a problem for the cognitive sciences, and ignore its broader application to empirical science per se.

of the topic, for instance by Van Fraassen (1977, 1980) and Kukla (1994), among many others.

A second reply is to offer further data to support one theory over the other(s), assuming that the successful addition of a new level of explanation will tip the scale in favour of one particular theory. One type of data could come from computations or simulations (cf. Bechtel and Abrahamsen, 2010). Another kind of data could come from neurological evidence, such as imaging data. The thought guiding both of these approaches is that getting a better perspective on what a human brain is computationally capable of in a given situation (in cases of computational evidence) or on what is actually going on in a human brain during a given task (in cases of gathering neurological data) will also eliminate some theory choices or support some over others.

Now, in order to answer this specific objection, I think I can show that the argument against EPCL actually is in a relevant sense different from the general underdetermination argument, and cannot be generalised in this way. Also, the two responses given above cannot help in resolving the problem, as of yet. While this latter point is not particularly encouraging, it suggests that the problems of developmental theories of concept learning are in some sense qualitatively different from those of other strands of psychology.

The reason why the argument against the EPCL hypothesis cannot be fully assimilated to the standard underdetermination argument is that the study of infant cognitive development and of infant concept learning is far more constrained in its methods than most other research areas of the cognitive sciences. Specifically, studies of adult cognition can rely on a wide range of imaging techniques, such as EEG, MEG, fMRI, and the like, which aren't available for the study of infant concept learning. This point holds even though some of these methods are starting to find applications in studies of early ontogeny. For instance, Andrew Meltzoff and his colleagues (Marshall and Meltzoff, 2011; Saby et al., 2012) did EEG studies with 15-month-old infants, testing their brain activity related to imitation behaviours. Another possible exception to my general point are eye-tracking studies, which are starting to be done with young children, in studies such as Hepach et al. (2012)'s investigation of two-year-olds' pupil dilation and sympathetic arousal in helping situations. Despite successes such as Meltzoff's work on imitation and Hepach's work on social cognition, the possibilities of using similar methods in typical concept learning studies seem to be rather limited because neither of these paradigms has yet been used in concept learning studies with prelinguistic infants. Neither Murphy (2002), Mandler (2004), nor Carey (2009) report the use of these methods, and I haven't been able to find other relevant instances of their use.

In conclusion, I take it that these two objections don't harm my argument against the EPCL hypothesis. So, with the problem set up, I want to go forward and discuss how it affects two prominent proposals for concept-learning mechanisms, Jean Mandler's and Susan Carey's.

3.3 Jean Mandler

Mandler (2004) offers a concept learning mechanism which she describes as independent of the HF model, and which she regards as the solution to Fodor's challenge. It is characterised by a minimal innate foundation sustaining the mechanism of Perceptual Meaning Analysis (*PMA*). *PMA* relies on four core theoretical claims:

- There is a fundamental difference between perceptual and conceptual representations. Perceptual representations are procedural (roughly, part of 'knowledge how'), whereas concepts are declarative (part of 'knowledge that'). This implies that only the latter, but not the former, can be used in personal-level thought.
- Concepts can be extracted from perceptual representations. A concept-learning mechanism based on experience uses information inherent in perceptual representations to create concepts from it. This implies reinterpreting perceptual information, for example by abstracting away from less salient information and only keeping a schematic representation of an event.
- The earliest concepts are redescriptions of image-schemas, as postulated by Lakoff and Johnson. Image-schemas are complex perceptual representations standing for aspects of a given spatial relation or event. Paradigmatic examples are 'contact' and 'path', which serve as underpinnings for conceptual representations like CAUSATION. The redescription process is driven by attention, which brings the image-schematic content into awareness, and can make it part of declarative knowledge (i.e., the conceptual system).
- The earliest concepts are global, higher-level ones such as ANIMAL or CONTAINER, rather than basic-level concepts such as DOG or CUP. If concepts stem from image-schemas, then they will be determined by some quite high-level properties (such as path of movement) rather than by secondary perceptual characteristics (such as colour patterns). Arriving at basic-level concepts happens through a process of differentiating the higher-level concepts into finer-grained ones that are more attuned to perceptual differences between members of the higher-order category.

First, Mandler strongly emphasises the importance of the distinction between perceptual and conceptual representations, and the corresponding distinction between perceptual and conceptual categories. At the foundation of this distinction and its relation to procedural and declarative knowledge lies the insight that perceptual representations are much richer in content than conceptual representations. In any given visual display, there are too many stimuli and details to consciously take in simultaneously, and we usually only focus on one part or aspect at a time - this means that we need attentional processes to limit the visual input we can process. At the same time, we still *see* the whole scene, so here, the content of visual experience and the content that is attended to and cognitively available seem to come apart. For Mandler, this is the primary reason why we should accept the distinction she makes.

Second, we see that Mandler still wants perceptual inputs to play a pivotal role in concept learning. For this to work, she needs to explain how the transfer from perception into cognition happens. The details on this will follow in the next section, where I present her concept-learning mechanism, Perceptual Meaning Analysis.

Third, Mandler subscribes to some of the very specific hypotheses cognitive linguists make about the foundations and the nature of human concepts, as for instance formulated by Lakoff (1987) and Johnson (1987, 1995). For our present investigation, the thing to note is the central claim of cognitive linguists regarding the origin of a large class of concepts: they claim that spatial concepts are foundational for a very wide range of human cognition and human language, which is evident in an abundance of spatial metaphors we use to think about the world and specifically about abstract objects (cf. Mandler, 2004, p. 78f.).[5] Mandler's commitment to image-schemas brings another noteworthy difference to Fodor's account of concepts (as reviewed in the previous chapter) into the spotlight: if a concept like CONTAINER is based on image-schematic representations, then it will be 'structured' while still being a lexical primitive. A schema of a container necessarily involves an inside, an outside and a boundary, or so the reasoning goes.

Fourth, we find that Mandler emphasises a non-standard order of the availability or acquisition of infants' first concepts. At least since Rosch and Mervis (1975)'s work on basic-level categories, it seemed certain that the developmentally most important categories were those of intermediate-level concepts (see also Mervis and Crisafi, 1982). Consider the example of cats: the concept CAT is the intermediary level between ANIMAL (the superordinate-level concept) and MINX, SIAMESE, and other cat variety concepts, which form the subordinate level of cat concepts. Within the developmental community, this fourth claim is the most controversial one, as the critical discussion in Murphy (2002, pp. 293–302) shows. Murphy contrasts Mandler's claim about the global level of infants' first concepts with the wide support for the case for 'basic-level' concepts, and shows that acceptance of the primacy of global concepts depends upon the acceptance of the perceptual-conceptual distinction (cf. Murphy, 2002, p. 295f.). Researchers who don't make the distinction, or who make it in a different way, could very well point towards results like Quinn et al. (1993) and argue that these early perceptual discriminations are the foundation for conceptualisation. According to Murphy, from this point of view, Mandler's sequential-touching evidence for global concepts can be reinterpreted as just showing a bias to touch some kinds of things more than others, and needn't reflect category judgments the child might make (cf. Murphy, 2002, p. 301). I will not engage in the debate, but note that I will revisit the proposal to tear

[5] For instance, thinking and speaking about time and about goals happens in spatial terms: time's passing, reaching a goal, and the like. I will not go into the many different image-schematic analyses cognitive linguists have proposed, and will assume Mandler's interpretation of image schemas is consistent and relevant to the current debate.

down the distinction 'perceptual/conceptual' later on.

However, Mandler's claim about the primacy of global concepts also has wider implications for her views on conceptual development: it implies that conceptual development is at its core a process of differentiation (cf. Niyogi and Snedeker, 2005, p. 4). Concepts like BIRD, DOG, or CAT are learnt by teasing apart the objects that fall under the concept ANIMAL along further dimensions that make them distinctive.

Now that we have given a broad stroke picture of Mandler's theoretical background, we are in a position to formulate the specifics of her concept learning mechanism, *PMA*, in short words.

3.3.1 Perceptual Meaning Analysis

According to Mandler, *PMA* works as follows:

1. Perception: The learner perceives an event P, which serves as the perceptual input p.

2. Analysis: The learner analyses p for salient features, and primarily spatial features such as movement or structure. The analysis uncovers an image-schema, or a combination of image-schemas, i. Only a small part of what is perceptually represented is used for the image-schematic interpretation of P. Linking i to p is the first result of *PMA*.

3. Redescription: Through a process of associative learning which includes attending to the occurrence of i in perceptual events like P, i gains access to conscious thought and thereby becomes a conceptual representation c.

4. The result: After this process, c is a conceptual representation, standing for the concept C. At the end of the process, C has been learnt through experiences of P.

It is useful to distinguish c and C because the former is a very basic kind of concept, and not the only conceptual representation of C that the learner could learn – at some point, she might learn a word for C, and other conditions beside the image-schematic ones expressed through c that are unique or defining for C.

The above is a general but focused representation of the mechanism, based on Mandler's descriptions of *PMA* in a variety of publications (Mandler and McDonough, 1996; Mandler, 2004; Niyogi and Snedeker, 2005; Mandler, 2008, 2010). Even before raising any objections below, I have to raise two cautious criticisms:

Meaning Mandler explicitly commits to 'meaning' as standing for 'concept', but also ascribes meanings to image-schemas, which she describes as non-conceptual (cf. Mandler, 2004, p. 67).

Analysis Mandler doesn't sufficiently specify what kinds of processes 'analysis' is supposed to involve. She introduces several candidates such as 'redescription' or 'abstraction', but doesn't give a detailed discussion of their specifics.

(Meaning) will turn out to be a problem later, while (Analysis) merely poses an interpretative challenge, at least in the narrative of this chapter. *PMA* crucially depends on the availability of image-schematic representations to create concepts. They comprise the bridge between rich, uninterpreted perceptual representations, and more narrowly constrained conceptual representations, which gain their cognitive worth exactly because they can be applied to a variety of different experiential encounters with exemplars of the concept (i.e. in various kinds of rich perceptual contexts).

As a hypothetical example, let's go through the process of acquiring the concept ANIMAL according to Mandler: young infants attend mostly to movement, and can distinguish the more irregular kind of movement of biological agents from a more rigid kind of inanimate motion early on (cf. Mandler, 2004, p. 71). This capacity requires a basic way of identifying objects as the kinds of things that move in (roughly) one unit. Third, they observe that some agents start moving by themselves, whereas others need to be put in motion. Mandler takes this to mean that they have the image-schemas:

> Self-starting Path (No Contact)
> Contingent interaction without Contact (Link)
> Rhythmic (biological) Motion. (Niyogi and Snedeker, 2005, p. 6)

In observing animals in visual scenes, the infant will focus their attention on the most salient elements of the scene, which would be moving objects. They analyse that the situation contains a biologically self-moving agent by applying these image-schemas. This enables them to redescribe the image-schematic representation of the animal into a consciously used representation, which will be the general concept ANIMAL. Mandler doesn't give a dedicated description to this 'redescription' process, but I think it is a process of association and pattern formation (cf. Mandler, 2004, p. 49) that elevates the original perceptual representations – the image-schemas, or combinations thereof – into the learner's awareness.

In the quasi-definitional form the concept will retain from its image-schematic origins, it is very useful for the infant because of its generality: it applies to a wide range of perceptually dissimilar objects (birds, dogs, ants, etc.), all of which are animals. The global nature of the first concepts, as stipulated by Mandler, certainly has its attractions because such concepts are powerful tools for categorising and understanding large classes of objects of experience.

3.3.2 Empirical evidence for PMA

Jean Mandler deserves a lot of credit for enriching the developmental literature with new experimental paradigms and with controversial findings regarding the primacy of global concepts before basic-level concepts. She describes the results of her studies in great detail in Mandler (2004), and explicitly endorses the view that the following techniques are the best kind of evidence for her concept learning mechanism. She has proposed two general sets of empirical evidence for *PMA*:

Sequential touching & object examination Young infants (up from 7 months) are shown to be able to group objects along categorical dimensions. This type of evidence supports the claim that global concepts are acquired first.

Deferred imitation & generalized imitation Young infants can reproduce certain imitated behaviours, some even generalising from the originally witnessed event, when they get the chance to perform the action. This supports the idea that they can generalise an image-schema-based concept to other instances where the schema is applicable.

As I have introduced these kinds of evidence in Section 3.2 above, I will focus on emphasising their relevance for *PMA*. The first group of studies is immediately relevant to the learning mechanism because it provides evidence for the importance of image-schemas in concept learning. This is so because image-schemas are bound to provide more general, global, kinds of concepts. If image-schemas are at the core of concept learning, then it is reasonable to assume that a process like Perceptual Meaning Analysis might operate with them.

The second group of studies is a bit less directly relevant to the learning mechanism, but it supports the broader framework Mandler uses. The infants' successes in imitation behaviours serve as evidence for the availability of global concepts like ANIMAL. Furthermore, it supports the view that the infants actually do have *concepts*, not just some kinds of subconceptual abilities, because the imitation behaviour requires the ability to make inductive inferences based on observation and on the classification of the available play materials. Inductive inferences, in turn, are a paradigm example for conceptual, cognitive abilities.

With Mandler's model established and with her evidence put into context, I want to continue by raising two problems for her approach.

3.3.3 Problems for PMA

While I regard Mandler's proposal as innovative and compelling, I have identified two main problems with it that will keep me from endorsing it as a successful reply to Fodor's challenge, and which I will detail in the following.

PMA can be reinterpreted as an instance of the HF model

My main objection to Mandler's *PMA* is that it fails to overcome Fodor's challenge because there are very compelling reasons to interpret *PMA* as an instance of the HF model. Both Fodor and Carey raise aspects of this problem that are hard to ignore, and difficult to overcome for Mandler without revising her theoretical commitments, as I shall argue now. Because of this, I propose a two-pronged approach, and will show how both Fodor's and Carey's point independently lead to the conclusion that *PMA* cannot explain concept learning without appeal to the HF model

Fodor (in Niyogi and Snedeker, 2005, p. 9) argues that, contrary to Mandler's own assertions, her model cannot work except as a version of the HF model. As we have seen above, *PMA* works by analysing the content of perceptual representation, using available image-schemas to interpret the contents of those perceptual representations. I would agree with a Fodorian interpretation of this – it is equivalent to constructing a hypothesis that is then confirmed by perceptual evidence. After all, there has to be a pairing of image-schemas – the carriers of at least some of the content the final concept is going to have – and perceptual representations if a stable image-schematic representation of an event is to be constructed. This is going to be the case even on a reading that succeeds in drawing a clear line between image-schemas and concepts, contra the (Meaning) worry above. Such a pairing can plausibly be regarded as the constructing and testing of a hypothesis, since a case of mismatches, or of pairing a perception with the wrong kind of image-schema, will have to prompt a different kind of image-schematic representation to be considered, i.e., a new hypothesis.

Consider the following as the main point of overlap between the HF model and *PMA*: for both, the concepts or image-schemas, qua materials for constructing a hypothesis, or qua comparison elements in *PMA* respectively, must already be available for being paired with a piece of (perceptual) evidence. So, they cannot be learnt simultaneously with the concept that is thought to be learnt in the process. Fodor uses the example of learning something new about dogs, for instance that they bark. Suppose a child already has the concept DOG and now observes a barking dog for the very first time. For a *PMA*-type explanation of the learning process for BARKING, we would need a set of image-schemas that can be used to represent the aspects of the scene that pertain to the barking. We can assume, for the moment, that there might also be auditory (image-)schemas. The relevant schemas will be assembled, and will undergo the redescription process, leaving the child with the new concept of barking. However, Fodor can immediately point out that this explanation directly fits his idea of HF model-type learning: to differentiate between two kinds, he claims, one surely needs a way of *thinking* about them as in a relevant way different from one another, such as grouping items along the distinction 'barks/doesn't bark'. To do this, one needs the concept BARKING, and not just the initial image-schematic concept ANIMAL. As there is no aspect of the learning episode that Perceptual Meaning Analysis can explain and the HF model cannot, it seems that there are no reasons not to regard *PMA* as an instance of Fodor's model.

Coming from a different angle, Carey (2002) voices a similar problem in her comments on *PMA*. Facing the problem of creating concepts, she asks:

> Where do the categories represented in the image-schematic meanings themselves come from? (Carey, 2002, p. 47)

The problem thus is that explaining concept learning through image-schemas opens the question of the origin of image-schemas. Are image-schemas learnt, or are they innate, for instance as properties or biases of the human sensory apparatus? And if image-schemas are supposed to be learnt, then how are they learnt? Even if we were to grant that the process that gives us concepts from image-schemas is a case of concept learning, which is doubtful in light of the above criticism, we have only shifted the problem down one level. Additionally, in light of the (Meaning) worry I raised above, there is still room for arguing that the entire step of getting concepts out of image-schemas is nothing more but unnecessary bells and whistles, which can plausibly be seen as a change in beliefs about the already available image-schema-concepts.

Mandler tries to escape this interpretation of her theory by downplaying the role of comparisons between members and non-members of a given category (cf. Mandler, 2004, p. 68). The thought behind this is that comparisons do involve an application of available concepts: in order to see whether a given moving object on the floor is a cat or not, a perceptual representation of the cat has to be compared with CAT; if it doesn't pass as an instance of CAT in the comparison, then it isn't a cat. Instead, Mandler wants to emphasise the transformative aspect of *PMA*, which turns perceptual representations into a different kind of representation (cf. Mandler, 2004, p. 70). The idea behind this is that *PMA* is capable of making a perceptual representation, for instance of a cat playing with a ball of wool, usable in forming thoughts about the cat and the wool, and that this has to involve transforming the perceptually given into something different.

However, this won't help with examples like learning the concept DOG as described above. The differentiation which she has in mind (dividing animals into dogs and non-dogs by some criterion available to conceptual thought) precisely requires a cognitively accessible difference; it requires a measure to determine whether something is a dog. Additionally, as alluded to above, the question of what is involved in transforming percepts into concepts is not sufficiently addressed in Mandler's work. Since the best reconstruction I can come up with involves a mapping of image-schemas to perceptual representations, the 'transformation' explanation brings us back into 'comparison' territory, as it were. The mapping that has to happen is a case of comparing a perceptual instance to a conceptual representation, which is an act of comparing the two. If my reconstruction is right, it doesn't transform the perceptual representation into a concept.

Mandler further tries to escape the problem Carey raises by conceding that some image-schemas might be innate (e.g. Niyogi and Snedeker, 2005, p. 6). This solves this problem in part, and in a way that Carey can easily accept (cf. Carey, 2002). I say that it is only a partly solution because conceding that some image-schemas could be innate does nothing towards explaining how the non-innate image-schemas are acquired. Even with such a concession to Carey, Fodor's criticism is still unaddressed, and *PMA* still in jeopardy. Taking together the two parts of the criticism I raised, it is difficult to avoid the following conclusion: if one removes the necessity of learning image-schemas, and the possibility that *PMA* could be anything else besides an instance of the HF model, then there remains nothing of the form of a CLM.

I submit that it is difficult for Mandler to escape this problem, unless she makes major changes to her theory, such as collapsing the perceptual/conceptual distinction. If she did so, she could potentially argue that perceptual representations can come to play a role in conceptual reasoning, which would escape this particular objection. This move would make her concept-learning mechanism potentially compatible to other, more perceptually based mechanisms, such as Lawrence Barsalou's simulator-based approach (Barsalou, 1999), or Robert Goldstone's Perceptual Learning approach (Goldstone, 1998), which I am going to discuss in the next chapter.

The evidence for PMA is not sufficient to sustain the theory

When considering the role Mandler's own experimental research plays for supporting *PMA*, I have found that the two kinds of experimental evidence that she brings to the table don't actually talk about the learning process. Admittedly, it can be taken as support for some of her theoretical commitments, like the existence of global-level concepts that might be based on image-schemas. Yet, this doesn't change the state of affairs regarding the relation between the proposed concept-learning mechanism and the evidence that is provided.

Furthermore, *PMA* and Mandler's experimental work fall prey to my argument against the EPCL hypothesis, as presented above. The intermediate data that she can produce, such as touching patterns, doesn't determine which kind of cognitive mechanism supports the alleged learning process. It falls squarely within steps 3-5 of the argument against the EPCL hypothesis. Granted, an infant will present a set of intermediary data points, such as touching times, which are listed in premise (4). These will count as amongst the strongest kind of evidence that the study can afford us with, pace premise (3). The problem with this is formulated in premise (5): the touching times are compatible with too many potential theories. This is exactly the underdetermination problem that I have raised above.

An additional aspect of the problem for Mandler's evidence is that neither of the touch- or imitation-based paradigms has a sensitivity to the temporal fine-grainedness of the concept learning process that is greater than the dishabituation studies'. Already from this point, it should become clear that the argument as given in Section 3.2 applies equally for Mandler and for Quinn.

To emphasise the problem *PMA* faces, let's again look at an example for a concept learning model that gives evidence for a cognitive change taking place during the learning process: the Wisconsin card sorting test, as discussed by Fodor (1981). A subject has to learn the referent of the new concept FLURG. Let's assume 'flurg' refers to GREEN OR TRIANGULAR. The subject gets a set of cards with geometric, coloured objects shown on them. After every card, she has to judge whether it shows an instance of FLURG or not, with the experimenter confirming or disconfirming each hypothesis. The test is carried out until the criterion is reached (e.g. 20 correct guesses in a row).

This setup is able to track something like learning because it relies on the learner's hypotheses: each new card requires her to state her hypothesis indirectly, by sorting the cards right ('yes, this is flurg'/'no, this isn't'). So, even though the experiment doesn't directly track mental goings-on, its behavioural results pull strongly in the direction of the HF model.

Contrast this again with Mandler's generalised imitation paradigm. One could do Mandler's task in a similar way, by measuring how often they make the right kind of generalisation about whether one should give an animal or an artefact a drink – 'this one drinks'/'this one doesn't drink'. But this by itself doesn't speak for Mandler's learning model: her explanation requires postulating that the infant has formed a new concept during the experiment by Perceptual Meaning Analysis. We have seen that *PMA* is at its most basic a process of obtaining conceptual content from perceptual content by applying image-schemas to the latter, which subtract less salient or less relevant aspects of the perceptual representation. This explanation builds a lot of extra assumptions into the process, which the hypothesis-formation model just doesn't need, such as the possibility of abstracting concepts from perception by mediation of image-schematic representations. Now – as I've argued above – if both models explain the data to a degree, one needs additional criteria to decide which explanation is better. Some such criteria would be theoretical simplicity, or parsimony, or good integration with additional data, or fitting analogy to other learning experiments. If we take some of these into account, I regard it as safe to say that the HF model is the better explanation until we have found more evidence that either supports Perceptual Meaning Analysis better than it does support the HF model, or which is even incompatible with the HF model. I haven't found this kind of evidence in Mandler's work.

I conclude that *PMA* is not supported well enough by her data from early ontogeny research, and that *PMA* cannot escape a reinterpretation as an instance of the HF model on the conceptual level.

3.4 Susan Carey

Susan Carey doesn't deny that the HF model is important and integral for concept learning, but proposes that accepting it doesn't lead to Fodor's paradox, since many cases of concept learning involve the enrichment of the conceptual system. In her book *The Origin of Concepts*, hereafter also *TOOC* (Carey, 2009), she calls her learning model Quinian Bootstrapping (*QB*). The Münchhausen metaphor of pulling yourself out of the swamp by your own bootstraps is used to symbolise the ascent into a conceptual system that is incommensurable to the initial one. The name also references W.V.O. Quine (Quine, 1953, 1960), whom Carey credits as the main proponent of bootstrapping as a learning mechanism – even though she disagrees with his Empiricist leanings. I will present the main theoretical commitments, the mechanism, and the main pieces of empirical evidence for *QB* in the following, and will then turn towards a set of potential problems for it.

3.4.1 Foundations of Quinian Bootstrapping

Quinian Bootstrapping relies on the following theoretical claims:

- Young children learn 'intuitive theories' that supplement their innate stock of concepts, which are part of Core Cognition. Core Cognition is a set of innate concepts and capacities that are at the foundation of all future developmental achievements. Intuitive theories are like scientific theories to a degree, but are more limited in scope and application.

- Like in the sciences, different stages of intuitive theories are incommensurable: for example, the adult concept of weight can not be expressed by the concepts young children have available – their understanding of what WEIGHT means is radically different from the adults'.

- Following Nersessian's 'cognitive-historical analysis' of theory change (Nersessian, 1992), the processes of theory construction by scientists and by children share many important features. This is most evident in cases of encountering new phenomena that the existing theory cannot explain. Then, new explanatory constructs – such as analogies or models, together with hypotheses about the relation between the construct and the available knowledge – are used for forming new representations of the concepts and laws that constrain and explain the phenomenon.

- Conceptual content is best understood by a dual-factor theory, which relies on narrow and wide content. Narrow content, or conceptual role, is determined by a concept's relations to other concepts, while wide content is the meaning a concept gets from its referent. Carey relies on this to sustain the idea that new concepts can be learnt in the theory-like way she proposes, going from minimal content (only narrow content) to full content (narrow and wide content fully established).

At the centre of the first claim, we can find two main ideas in Carey's work: a restricted version of Concept Nativism and a commitment to a version of the Theory-Theory of concepts. Carey's Nativism is based on the idea that there is an innate set of concepts and psychological mechanisms that form the core of human cognition, and which in large parts stays online throughout the human lifespan; hence the name 'Core Cognition'. In *TOOC*, she devotes several chapters to main aspects of Core Cognition, such as the concepts pertaining to objects, numbers, and agency. Only after having established her point about the innate parts of human cognition does she address the question of concept learning. Since I don't aim to discuss the topics surrounding Concept Nativism, I will leave this part of Carey's theory largely unaddressed and will turn to the second main point regarding the Theory-Theory (*TT*).

Carey uses a specific version of the Theory-Theory of concepts, which I propose to call 'developmental Theory-Theory'. The developmental Theory-Theory needs to be differentiated from several other types of *TT*. As Machery (2009, p. 100f.) observes, there are two origins to *TT*, one in categorisation studies (Murphy and Medin, 1985) and one in developmental research (Carey, 1985; Keil, 1989). Furthermore, as Machery (2009) analyses, some theorists claim that *TT* implies that concepts are theories themselves, whereas others only hold that concepts are parts of theories. Carey's developmental version of *TT* belongs in the latter category, and avoids some of the former position's problems.

Apart from the idea that children learn intuitive theories, as introduced above, there are two further claims characterising Carey's version of the Theory-Theory. First, paraphrasing Carey (1985, p. 198), concepts have to be seen and evaluated within the framework of a theory they are a part of. Second, in line with researchers like Gopnik and Meltzoff (1998), Carey holds that young children also use scientific methods when learning their intuitive theories. I will go into more detail on this point further below.

The second claim about the incommensurability of intuitive theories has to be viewed from the background of Carey's developmental concerns: if concept learning is supposed to create a set of concepts with greater expressive power than its predecessors, then some of the learnt concepts must express something the earlier systems couldn't express. So, in this sense, two systems of concepts can be seen as incommensurable.[6] In this generic sense, simpler and more abstract, or theoretical, kinds of incommensurability are possible. Among the simpler kinds, one can imagine a learner who initially has no concepts related to weather phenomena, but learns concepts like RAIN, SNOW, and SUNSHINE at some point. Here, we have a case in which a system of concepts has been enriched. Another case from theory learning is learning the difference between heat and temperature. It is conceivable that somebody uses both concepts interchangeably, with both meaning something like FELT AMBIANCE WARMTH. The distinction is lost on somebody with that unified concept. After she learnt that HEAT refers to the amount of thermal energy in a given system, and that TEMPERATURE is a measurement of that thermal energy, she can make informed judgments about topics like heat transfer and temperature changes, and how they are related.

The third main claim ties in with the previous two, insofar as it stresses the analogy between the two cases of concept acquisition and theory change. It further specifies the methods involved in the kinds of concept learning that Carey investigates. Nersessian (1992) discusses these methods as part of her cognitive-historical analysis of scientific theory change, which Carey adopts as examples of Quinian bootstrapping. I will return to Nersessian's work below, when considering the empirical evidence for *QB*.

[6] When talking about incommensurability in the following, I will use it in this sense. By this, I aim to avoid debates about other uses of the term, such as the multifaceted use in Kuhn (1962).

Finally, relying on a dual-factor theory of meaning, as advocated by Ned Block (1987, 1998), guarantees the success of *QB* as a concept-learning mechanism, and ties in very well with the developmental Theory-Theory. Put very succinctly, dual-factor theories acknowledge that both external factors, such as referents in the external world, and internal factors, such as a concept's role within a larger conceptual framework, constitute a concept's meaning. By this, these theories can avoid some problems of content internalism and externalism alike.[7]

With this framework set up, we can easily reconstruct concept learning by Quinian Bootstrapping.

3.4.2 Analysis of Quinian Bootstrapping

According to Carey, *QB* works as follows:

1. The learner encounters a new phenomenon P that she doesn't understand yet/ lacks the concepts for.

2. By learning a set of initially content-less placeholder concepts (words, symbols), the learner constructs the vehicles that will come to be the concepts related to P. They are at least by a minimal standard defined in relation to one another and only partly interpreted by reference to previously available concepts.

3. Hypothesis formation and testing come into the picture at this point, where a model for understanding the target domain is constructed. Modelling techniques are essential here, as they provide the foil for making the transfer of knowledge about the target domain into the placeholder structure.

4. At the end of the learning process, the learner can fully replace the model with the actual wide conceptual content, i.e., the extensions of the concepts.

I want to show how Carey's theoretical claims from above relate to *QB* as a concept-learning mechanism, and illustrate it with an example from *TOOC*. In several recent discussions of Carey's work (e.g. Shea, 2011; Keeler, 2012), the illustrating example has been the learning of natural numbers. However, since this is a topic of considerable and voluminous debate in psychology, I want to use the example of the concepts MATTER and WEIGHT. This will help in keeping complexities extraneous to the learning mechanism to a minimum.[8]

To facilitate matters, I want to introduce a convention Carey uses in *TOOC* for talking about the differences between children's and adults' concepts of physics: she calls the child's initial theory of physics PT1, and calls the adult system PT2. At the

[7] Like all the other foundational claims of Carey's model, this one will also go unchallenged for now, and without a deeper embedding into the debates about meaning, reference, content, etc.

[8] Carey (2009) discusses the example in great detail, which I cannot reproduce in the present piece of work. I will therefore fictionalise the process in the following, but will return to some of the details in Section 3.4.3.

beginning of the learning process, children have an undifferentiated PT1 concept of WEIGHT/DENSITY. Carey discusses this with an example from Piaget's work (Carey, 2009, p. 385ff.): if asked whether a single piece of popcorn would be heavier than the piece of corn it was made of, children would answer in the affirmative. The implicit assumption is that a larger object will be heavier, ignoring changes in density that happen when a piece of corn pops. Carey argues that the children's concept of weight is indeed incommensurable to the adult concept since the adult's statements about weight just don't make sense when one applies the child's concept.

The learning process can begin at this point, for instance by showing the learner an inconsistency in their judgments. This could be achieved by putting the piece of popcorn on a scale, taking note of the number it displays, and then repeating this with the piece of corn. As both will show the same number, the learner will see her original verdict put into question. This is an example of step 1 above.

In the second step, the learner has to take notice of the notions at the core of the part of PT2 which are related to the concepts WEIGHT and MATTER. She might for instance have a teacher explain the relation between weight, mass, and density, and will have to learn about the connection between mass and matter. In relation to this, she will also need to acquire the distinction between material and immaterial things. This list is just a snippet of the complete physical theory of matter, but let me assume for the moment that it will be sufficient. At this point, the importance of the Theory-Theory of concepts for Carey's project should become apparent. The concepts I just introduced all stand in close relations, and a complete understanding of any of them requires an understanding of the others as well. Furthermore, another major point becomes immediately important: just by learning the names for the new concepts of PT2, the learner doesn't yet learn the complete concepts. The names serve as placeholders for the concepts to be learnt in the following steps. Even after having memorised them all, and knowing the relations between them, the learner would still be likely to, for instance, be convinced that a very small piece of styrofoam has no weight (cf. Carey, 2009, p. 405).

So, it is evident that the dual-factor theory of conceptual content plays a central role right from the outset of the learning process. It guarantees a central way in which Carey can overcome Fodor's challenge because it divides several parts of the conceptual content. Acquiring a set of placeholders is unlike hypothesis formation because it is at its most basic just a process of memorising symbols which aren't yet in contact with any kind of referent. At this stage, the only part of the concept's content that needs to be fixed is its relations to a set of other concepts – its conceptual role.

It is however possible for the new concepts to be connected to the previously available set of concepts. Such a connection might very well be temporary if the learnt concept comes to be incompatible with the prior concept. A learner of the PT2 concept WEIGHT could try to apply their PT1 concept in trying to understand a given phenomenon, but it might easily mislead her, as in the examples above. The old concept of weight is still useful, since it grounds the to-be-acquired concepts in a given part of the learner's experience. By contrast, we can imagine a case in which all new concepts are completely devoid of contact to our previous experience, such as learning a new language without overlap with our own language, and with a radically different kind of structure.

After having set up the placeholder symbols for the new concepts, the third step consists in applying a set of scientific methods and problem solving heuristics in order to relate the symbols to the right kinds of referents – the wide conceptual contents. In cases such as WEIGHT and MATTER, the referents arguably still aren't exactly tangible objects, but abstract ideas. Yet, these ideas apply to natural phenomena, and thus I think it is reasonable to stipulate that a complete PT2 concept of weight will be one which can be correctly applied to observed phenomena. The methods for establishing this connection are manifold. Carey relies on the four methods that Nersessian presents as the main instruments of scientific theory change: analogy, imagistic reasoning, thought experiments, and limiting case analysis.

In the case of learning WEIGHT, Carey describes a string of methods that will help a majority of children to learn the PT2 concept. In this, she follows Smith (2007)'s work with junior high school students. First, in order to learn the "mapping of weight to number" (Carey, 2009, p. 440) that is required, students learn to measure weights and volumes. Next, students have to extend their measurements by applying them to objects that they cannot weigh with their instruments, such as very small objects. These steps are necessary to grasp "weight conceptualized [as] a continuous, extensive, property of all material entities" (Carey, 2009, p. 441). Next, to clarify the relations between weight, density, and volume, the students use a visual analogy: drawings of geometrical shapes filled with a given amount of black dots represent volumes with a given density – the more dots there are in a figure, the heavier and denser the represented object. To facilitate the learning, other examples are used to put the visual analogy in context, such as putting sugar in a drink (increasing the density, and thereby the weight). Furthermore, during these hands-on kinds of exercises, the students also use the formulas that they have learnt at the beginning – the original placeholders. The interplay of different methods and representations of the subject matter helps a majority of students in learning the PT2 concept of weight. The aim is that the learner will have an understanding of the new concept that is independent of the models and bridging conceptions that she used in the learning process.

Before turning to the evidence for *QB* as a learning mechanism, I want to raise a final point about Carey's proposal. Since *QB* relies on the capacity to use symbols as mental placeholders for full conceptual contents, it is developmentally constrained in that prelinguistic infants cannot learn by this kind of bootstrapping process (cf. Carey, 2009, p. 329f.). The earliest instance of *QB* at work in children is learning natural numbers, starting around 2 years of age (cf. Le Corre and Carey, 2007, p. 398). All earlier concept learning will have to be explained by another mechanism. I will return to this point in Section 3.4.4 when discussing the applicability of the argument against the EPCL hypothesis.

3.4.3 Evidence for Quinian Bootstrapping

The main kind of empirical evidence for *QB* as a concept-learning mechanism comes from two sources:

History of Science Drawing on Nancy Nersessian's work, Carey uses case studies reconstructing the cognitive processes behind the theoretical breakthroughs of Darwin, Kepler and Maxwell. These are evidence for the claim that science often progresses through partial changes in theories, and for the claim that modelling techniques are at the core of conceptual change.

Science Education Carey shows that successful models of teaching science in schools use processes that are part of *QB*. Examples include Smith (2007)'s model for teaching the concepts WEIGHT and MATTER. This serves as evidence that scientific theory-change is a close model for children's concept learning.

Nersessian (1992) provides much of the backdrop for Carey's learning model. While building on the groundbreaking work of Kuhn (1962), she also criticises it for adhering to a misleading psychological doctrine in emphasising the 'gestalt switch' as the main point at which a scientific concept gains a new meaning in a revolutionary new theory (cf. Nersessian, 1992, p. 5f.). Cognitive-historical analyses serve as evidence for *QB* because they uncover and track the thought processes and guides towards understanding that scientists have used to develop new theories, and to revolutionise their fields of study. Scientists' notes and publications can provide insight into the process of change.

Second, Carey uses Smith (2007)'s work as a model for applying *QB* in physics education. She argues that the successes of curricula based on learning through *QB* are more effective than the available alternatives.

3.4.4 Problems for Quinian Bootstrapping

While there are several other possible angles of criticising *QB* and its theoretical foundations, I want to focus on the two following ones, which specifically address issues regarding *QB*'s qualities as a concept-learning mechanism.[9]

Explaining the leap

First, I want to raise the question whether bootstrapping as a mechanism is a feasible explanation of conceptual change. Matthew Keeler (2012) suggests that the process by

[9] I will omit criticism of the Nativist component of Carey's theory, as well as raising problems for the dual-factor account of conceptual content, just to give two examples that would immediately come to mind.

which the actual new concept is established in a bootstrapping episode is not cognitively explicable, but more of a circumstantial event.[10] If the crucial step between setting up a placeholder structure and between linking these placeholders to their referents (in terms of wide content) cannot be explained except in terms of a 'leap' or of serendipity, then Carey faces a dilemma. A Fodorian could either argue that the leap is a brute-causal event that is best understood in terms of a triggering explanation, or he could argue that the 'leap' is just a disguised version of an inductive inference, which makes the learning an instance of the HF model. In either case, *QB* wouldn't capture a learning process, and wouldn't count as a solution to Fodor's paradox. I will spell both horns of the dilemma out and then offer a reply which defuses the worry.

At first thought, this objection poses a serious difficulty for Carey. After all, it renders her learning mechanism vulnerable to Fodor's brute-causal reinterpretation if she insists on the 'luck aspect' of Quinian Bootstrapping. A Fodorian could contend that any kind of leap from a partly interpreted concept to a fully interpreted, learnt, one is analogous to a case of acquiring a new concept through an unexpected blow to the head. The analogy in question is that both cases fail to give a cognitive explanation of the origin of the new conceptual content. Neither a concept acquired by a leap nor one acquired by blunt trauma displays any discernible content transfer happening in the acquisition process. To link this to Fodor's definition of what concept learning is, these two cases are causal mechanisms but fail to also be rational mechanisms.

If, on the other hand, Carey were to offer a more detailed explanation of the way leaps to new concepts are inductive or abductive, then one might just regard *QB* as a very elaborate instance of the HF model, and thus not a solution to Fodor's paradox. After all, inductive generalisations are at the heart of the HF model, and require the availability of the concepts-to-be-learnt. Similarly, for abductive reasoning, especially of the IBE (inference to the best explanation) kind, or of the original Peirceian kind (Niiniluoto, 1999, p. S437), one needs available hypotheses phrased in the concepts one wants to confirm or disconfirm. In order to show how an 'inductive/abductive leap' is different from a standard inductive/abductive argument, Carey would have to clarify what she regards as the nature of the leap involved in these processes. Since, as Keeler points out, she isn't offering any such analysis, the objection weighs heavily on Quinian Bootstrapping.

Still, Carey has a few possible ways of replying to this charge. The most promising one I can propose here is following Nersessian's observation that

> limiting scientific method to the construction of inductive or deductive arguments has needlessly blocked our ability to make sense of many of the actual constructive practices of scientists. (Nersessian, 1992, p. 13)

Nersessian makes this point in connection with the varieties of modelling techniques scientists have used to build new theories, such as "(1) analogical reasoning, (2) imag-

[10] Keeler made this argument in a talk at the 2012 Rudolf Carnap Lectures, at Ruhr University Bochum, and in an unpublished manuscript (Keeler, 2012).

istic reasoning, (3) thought experiment, and (4) limiting case analysis." (Nersessian, 1992, p. 12) For Carey to supplement her view, she could argue that the debatable 'leap' doesn't actually take place as an independent step, but is a part of the successful employment of these modelling techniques. As a hypothetical example, suppose that a learner of elementary physics familiarises herself with the analogy of electrical current to waterways, and comes to accept that electrons flow in a way that is similar to the way water flows down a river. By accepting this auxiliary assumption, she might have taken the last step to a complete and (provisionally) correct concept ELECTRON. In this example, there is no leap at all, just a final step of accepting the adequacy or accuracy of an analogy, which becomes a go-to resource in thinking about electrons. A similar case could be made about the other modelling techniques Nersessian proposes, but I will leave it at this point for now. The conclusion I want to draw from these replies to the 'leap objection' is that there are potential ways of avoiding the necessity of explaining the leap. The dilemma that the Fodorian suspects arises from ignoring the genuinely cognitive (against the first horn) and dynamic, modelling-based (against the second horn) processes involved in theory learning.

Evidence

Second, I want to confront the problem raised at the beginning and see if it threatens Carey: does her experimental evidence support her theory in a sufficient way? I submit that it does, since it appeals to reportable factors of the concept-learning process, like the communicative and the scientific methodology needed for creating a new conceptual system. Specifically, evidence for QB is not threatened by the argument against the EPCL hypothesis because it avoids its premise (4) about the limits of in-between data in infant concept learning studies. Her evidence doesn't fall under any of the descriptions in (4), and thus the argument loses its grip on QB.

Compared to the prelinguistic evidence that (4) speaks to, cases like following a young child's learning natural numbers, or of a high-school student learning the difference between heat and temperature are much richer in the data that they provide. One can follow increments in improving understanding by following verbal reports, just to give an example.

Also, as an added point, children and young adults, who learn by QB, can in principle participate in a much wider range of experimental paradigms than prelinguistic infants. This means that there will be a wider range of data that can potentially point towards one theoretical explanation of the learning processes rather than another.

Carey's views on prelinguistic concepts as such are not threatened either, since she doesn't give an explanation in terms of concept learning, but in terms of innate Core Cognition. However, besides the possibility of refusing to accept a Nativist proposal of this kind, her reliance on Core Cognition brings another serious worry: unless we find much earlier, prelinguistic instances of concept learning that can be explained by QB, we could be committed to a position that excludes the possibility of prelinguistic concept learning. Relying on a set of innate concepts, recognition/triggering mechanisms and maturation might be our only option to explain cognitive development before learning by Quinian Bootstrapping, should *QB* turn out to be the only feasible concept-learning mechanism. This is a serious lacuna for an account of concept learning in early ontogeny, and should not be taken lightly.

However, there might yet be other mechanisms that could in principle be compatible with Carey's ideas on Core Cognition and Quinian Bootstrapping, which we just haven't taken into account in this investigation yet. I submit that the successes of *QB* shouldn't yet discourage the search for mechanisms of prelinguistic concept learning, just as the problems of the EPCL hypothesis shouldn't yet discourage us from finding better ways of studying infant concept learning.

Insofar as the problem generalises to be an instance of the underdetermination of theories by data, there could potentially be a different version of this argument speaking against Carey's proposal, but I won't go into this possibility now. Besides, as I have argued above, the argument against EPCL is specifically a problem for research in early ontogeny, so any further argument concerning the underdetermination question should be regarded as completely independent if it focuses on other areas of developmental research.

3.5 Summary

I want to conclude that psychological research in early ontogeny faces a deep problem with regard to concept learning since the evidence for *prelinguistic* cognitive capacities doesn't sufficiently support the elaborate theories that are needed for a model of concept learning. The challenge for developmental psychology remains to find experimental paradigms that univocally support one learning model over another. Carey escapes this problem because she doesn't offer a prelinguistic concept-learning mechanism, but relies on nativist assumptions to explain prelinguistic cognition. While this is a debatable move on Carey's part, I choose to move on, and to accept *QB* as a potentially viable concept-learning mechanism while regarding Mandler's *PMA* as defeated by Fodor's challenge, at least until further notice.

Chapter 4

Perceptual Learning as a concept-learning mechanism

4.1 Perceptual Learning and feature-based approaches to concepts

Cognitive psychology has recently seen the development of several new models positing a perceptual basis for conceptual systems. This panoply ranges from proposals to eliminate the distinction between concepts and percepts altogether (Barsalou, 1999) to more modest appraisals of the relation between the two (Goldstone and Barsalou, 1998) to proposals for the creation of cognitive processes through experience with perceptual stimuli (Schyns and Rodet, 1997). As an important contribution to this line of research, Robert Goldstone's perceptual learning approach stands out and shall be the centre of our present investigation into the links between the perceptual and the conceptual.

Among the specific questions related to perception and learning, Goldstone and his colleagues and collaborators discuss the possibility and mechanisms of perceptual learning (Goldstone, 1998), the influence of perception on categorisation (Landy and Goldstone, 2005), the role of features of objects in categorisation (Schyns et al., 1998), and learning in early ontogeny (Goldstone et al., 2011). Their target is the more conservative fixed-feature approach, which holds that new concepts are constructed by using pre-existent, cognitively fixed features. Accepting Jerry Fodor's argument against concept learning can be seen as necessitating a fixed-feature approach to concepts, as I will show in the next section. Taking this conclusion as an undesirable outcome for any theorist who wants to maintain a notion of genuine concept learning, one might ask the following question. Assuming that cognition is at least a partly computational process, is there any way of having new symbols from perceptual origins enter the internal symbol system? As Goldstone answers this to the affirmative, I will discuss his proposal and point out two problems with it that need more consideration, before offering two alternative ways of my own for responding to Fodor's Challenge that are based on Goldstone's Perceptual Learning (henceforth *PL*) approach.

4.2 Fixed-feature approaches and Fodor's paradox

In order to see where Goldstone and colleagues aim when they criticise the fixed-feature position, I will first introduce the fixed-feature approach and the challenge it poses to research in perceptual learning. As we have seen in Chapter 2, Fodor goes through a number of steps to reach the conclusion that concepts, and by that also features, cannot be learnt.

First, Fodor construes learning mechanisms as rational-causal processes. He further notes that concepts that are not learnt are either innate or acquired in some other, non-rational way. Fodor also relies on a Rationalist premise, which posits that the primary use of concept lies in forming beliefs, as opposed to guiding actions or categorising perceptual experiences. Fodor argues that the only available, empirically tested model for learning is the HF model: learning a concept consists in forming hypotheses about it and testing them against the evidence. Thus, learning is a process of inductive inference. Now, what is already used for hypothesis formation is not learnt in the application (confirmation or disconfirmation) of the hypothesis. Thus, the concept must already be available to form the hypothesis, and so the concept was not learnt.

Fodor's conclusion is that all concepts are either innate or non-rationally acquired. This conclusion is supposed to affect theorising about learning concepts in all areas of the cognitive sciences, from developmental psychology to artificial intelligence research, since it affects the theory choices one has to make to explain the phenomena of these disciplines. Consider artificial intelligence (AI) – an important research aim of AI is to develop systems with human-like intelligence – computer programs that play chess like grandmasters, robots that move like biological organisms, and the like. The search for a theoretical foundation for such systems led to the proposal of the Computational Theory of Mind (CTM) – roughly, that the human mind can be best described as a system that works like a computer operating on symbolic representations. For a Fodorian computationalist, the number of symbols would be predetermined by the system, and so the symbols would be innate. Given the additional constraint that each symbol of such a computational mind equals one concept, one has arrived at a point where Fodor's paradox and the computationalist programme tie in. Landy and Goldstone (2005) describe such conceptions of CTM as being essentially linked to the idea that a fixed store of primitive, basic symbols is sufficient for successful cognition, and continue by saying that this classical version of CTM

> entails a fixed set of primitives, or at least demands that any alterations to the primitive set are not cognitively interesting acts. (Landy and Goldstone, 2005, p. 346)

Thus, we have characterised one stance towards Fodor's paradox, the fixed-features approach: it accepts the conclusion of Fodor's paradox, embraces the Radical Concept Nativism that it allows for, and denies any effect on the cognitive system that would count as learning a new primitive concept. By challenging the sufficiency of a fixed set of symbols for explaining human cognition, and by denying that changes to primitive symbols aren't cognitively interesting, Landy and Goldstone set out their alternative to the Fodorian position. They have developed a theory that we will presently take into account as a reply to Fodor's Challenge, as formulated at the end of Chapter 2. Landy and Goldstone propose the creation of cognitive symbols from perceptual materials, and they want to argue for the possibility of manipulating "systems of high-level categories" (Landy and Goldstone, 2005, p. 346) to better fit the demands of the cogniser. The question motivating the present investigation is: can Goldstone's theory of perceptual learning, and especially Landy and Goldstone's stance against fixed-feature languages, stand against Fodor's Challenge, and can it give a mechanistically and computationally credible account of human concept learning?

4.3 Perceptual Learning as a reply to Fodor's Challenge

The question before us is whether it is possible to enrich a symbol system through the manipulation or introduction of perceptual information, or perceptual symbols. Learning features, like other forms of concept learning, can, in an important sense, be seen to hinge on the possibility of arriving at thoughts one wasn't able to hold or express before. If we want to avoid Fodor's anti-learning conclusion, we have to provide an alternative empirical model for concept learning that cannot be reinterpreted as an instance of the hypothesis-formation-and-testing model without significant explanatory loss. Fodor denies that it is possible to provide such an alternative, whereas several recent contributors to the debate have tried to develop cognitive-psychological models that hold this promise.

In addition to their opposition to the 4th premise of Fodor's paradox, Landy and Goldstone also challenge Fodor's assumption that the primary use of concepts is in forming thoughts, as opposed to using concepts in dealing with the world – as in reacting to (sensory) inputs and producing (behavioural) outputs.[1] Grounding concept use and concept learning in perception does however not preclude the use of new perceptual concepts in higher cognitive activities – this is an important point made by Landy and Goldstone, e.g. in their discussion of changes in scientific reasoning through perceptual changes.[2] It is worth dwelling on this aspect of Goldstone's theory before turning to the core of Landy and Goldstone (2005)'s proposal. Goldstone investigates the possibility and the mechanisms of what he calls Perceptual Learning, following Gibson (1963):

> Any relatively permanent and consistent change in the perception of a stimulus array, following practice or experience with this array, will be considered as perceptual learning. (Gibson, 1963, p. 29)

On this definition, PL is a sensory as well as a cognitive process. For example, the changes in focus, or in attentional centre, in seeing something are at the same time changes in the categories pertaining to the perceived object. The repeated sensory contact with a certain class of objects will bring about a change in the way one thinks about these objects, which will in turn influence its perception, i.e. the sensory processes. Goldstone explicitly wants to trace the ties between these perceptual changes and the possible conceptual changes that accompany them. He holds that one traditionally neglected aspect of the relation between perception and conception is the influence that the conceptual system has on perception. In categorical perception, the learnt categories influence the performance in perceptual tasks. Especially in the sciences,

[1] This is not to say that sensory inputs and behavioural outputs play no role at all for Fodor. What he argues for in *LOT2*, however, is that the importance of having concepts lies in forming thoughts, and not in synchronising oneself with the world. Forming thoughts that relate to the world still requires them to relate to the world in the right way, hence the doorknob/DOORKNOB problem (cf. Fodor, 1998, and Chapter 2).

[2] Landy and Goldstone (2005) discuss changes in ontology, cognitive properties of groups of scientists, and changes in scientific practice through new perceptual capacities as cases in point.

there are multiple examples for this. Mathematicians can name several properties of a function just by looking at its graph. Similarly, after studying the geological categories and training to differentiate various stone samples, geologists have a sharper grasp of the differences between stone types, and are able to recognise them much faster than any layperson could (Goldstone et al., 2012; Goldstone and Hendrickson, 2010). This is also the second point Goldstone and Barsalou (1998) stress:

> [P]erception's usefulness in grounding concepts comes from several sources. First, perception provides a wealth of information to guide conceptualization. Second, perceptual processes themselves can change as a result of concept development and use. Third, many of the constraints manifested by our perceptual systems are also found in our conceptual systems. (Goldstone and Barsalou, 1998, p. 232)

The first statement of this quote, that our perception can be a source of information for our conceptual system, doesn't sound very controversial since it isn't very specific. In what way does perception inform conception? Even on Fodor's account, perception informs conception in so far as perceiving an object x can cause the triggering of the accompanying concept X. For Goldstone, and especially for Barsalou, there needs to be a more detailed description of the way in which perceptual information touches upon our concepts; a description which probably even does away with the distinction between perception and conception (cf. Barsalou, 1999). The second point is that changes in our concepts can influence our perception of the world. This bears on the present question in so far as it is the converse of the claim that Landy and Goldstone put forward to challenge Fodor, which is that perception can change our conceptual repertoire. If both of these directions of influence were part of the workings of the human mind, then the strongly computationalist position would either lose a lot of its plausibility, or would have to be reformulated to accommodate these interrelations. This is because strong computationalism is based on the fixed-feature approach, which denies that perception can change the original set of primitive concepts. Such an accommodation would however run against the self-proclaimed Rationalist position that Fodor adopts. Finally, the third point is especially important for Barsalou's project, but beyond the scope of the current investigation.

With these preliminaries set out, let's see how they form a framework for Landy and Goldstone (2005) answer to Fodor's challenge. Here is a short characterisation of Landy and Goldstone's main argumentative aims:

1. In learning about things we don't already understand, our cognitive system constructs specialised variable-feature languages that deal with these novel things.
2. The vocabulary of these languages consists of stimuli that we perceptually pick up and group as belonging to features, or feature dimensions.
3. New features can be learnt by applying the grouping mechanisms of unitisation and differentiation, the main players among other perceptual mechanisms.

As we have seen in Chapter 1, concepts are considered to be central components of human thought. Thus, they are rather highly developed parts of our mental lives, and conceptual thought is the epitome of cognitive activity. Now, many things that we think about are specific to a problem domain, like choosing a move in a chess game; others are central to modern human activities, like deciding which way to go to reach the nearest restaurant. Keeping with the computational tradition in the study of cognition, one can speak of different 'vocabularies' or symbol stores for different tasks, with some being used for a more diverse range of activities than others.

Landy and Goldstone follow this lead and frame the debate about concepts as one about the languages of cognitive systems. This is a common move, given that Computationalism treats cognition as symbol-manipulation, and a number of symbols, combined with some of the operations over these symbols, can with some right be called a 'language'. In the context of this chapter, I propose to call such a language a Language$_C$ (computational language), to highlight that the sense of 'language' is restricted as compared to that of a spoken language. Computationalists like Newell and Simon (1976) or Fodor and Pylyshyn (1981, 1988) favour a fixed Language$_C$, whose symbols are inherent in the cognitive system and sufficient for any kind of cognitive activity within that system – there is no need to import new symbols, since the given stock is supposed to express any proposition that the system would need to process. Biederman (1987)'s geon model is another example of a fixed Language$_C$, with the added twist that he attempts to posit perceptual representations – representations of basic geometrical forms – as a part of the innate stock of symbols.

To counter this model, Landy and Goldstone present what they call a "variable-feature language" (Landy and Goldstone, 2005, p. 347): a Language$_C$ that can be enriched with new primitive symbols, if new perceptual tasks require this. In Landy and Goldstone (2005), they characterise these enrichments as additions to particular sets of symbols, constrained by the category, or task, they are used for. In this, they follow Schyns and Murphy (1994)'s major contribution to feature-based approaches to concept learning.[3] This leads to changes in highly specialised vocabularies, and needn't necessarily affect the foundations of the Language$_C$. Landy and Goldstone talk of special-purpose Languages$_C$ and general-purpose Languages$_C$ in the cognitive system. While the latter aren't excluded by Landy and Goldstone as innate given that they are ubiquitous in the most basic cognitive functions, the former need to be learnt. This is because the tasks that they are needed for are highly specialised in one way or another: examples that their paper discusses include fine perceptual discriminations such as discriminating brightness and saturation, and scientific theorising and theory construction. Landy and Goldstone (2005, p. 348) compare the cognitive symbol system to LEGO blocks: some objects can only be constructed in a very cumbersome

[3] This has been pointed out to me by a reviewer of Stöckle-Schobel (2012). The functionality principle, that functional demands shape the perceptual processes of categorising new stimuli and forming new featural discriminations, from Schyns and Murphy (1994), has been a cornerstone of recent work in this area.

manner if one is restricted to using the standard blocks (think of sails for a pirate's ship), so adding LEGO sails to their repertoire facilitates that specific kind of building process. The disadvantage of these special parts, however, is that they cannot serve for much else except their originally intended function. This, again, echoes the constraints on special purpose Languages$_C$: the concept COLOUR SATURATION only has a very limited set of tasks for which it is needed, whereas the concept NOT has a scope that's equivalent to the generic LEGO blocks.

Now, in terms of the mechanisms of learning, Landy and Goldstone's theory's main component is the addition of a cognitive device at the interface of perception and conception, which slowly builds 'cognitive symbols' out of perceptual stimuli. By adding these new symbols to the symbolic building blocks of thought, this device is the agent of concept learning and conceptual change. The main operations in this system are unitisation and differentiation, two mechanisms which either unite previously separated conceptual elements, or split a vaguely bounded class into finer groupings. In my present investigation, I will focus on these two mechanisms, since they are central to the argument by Landy and Goldstone (2005). When linking their theories with other, related work in the field, like Goldstone et al. (2011), or Goldstone and Landy (2010), they introduce other ways of learning. These include processes that Fodor would classify as brute-causal acquisition rather than genuine learning, to which we will turn later in this chapter.

4.3.1 Unitisation

Unitisation can be described as a process of grouping several previously independent categories under one heading:

> When elements co-vary together and their co-occurrence predicts an important categorization, the elements tend to be unitized. (Landy and Goldstone, 2005, p. 350)

Here is an example of a process of unitisation learning: suppose you learn what a cup is by seeing various different cups and not-quite-cuplike objects. Something qualifies as a cup if it consists of a cylindrical container and a handle to the side of the container. The contrast class of cup-like objects consists of other configurations of containers and handles, like a handle spanning the top or the bottom of the container (the former looking a bit like a bucket), or with a handle only connecting with the container at one point (looking like a horn attached to the cylinder), or even just unconnected handles and containers. The rules of unitisation would incline you to unite the two featural elements (cylinder and handle) into a token of the concept if and only if they are in the right spatial configuration (handle on the side, both parts properly connected). Unitisation allows you to conceive of the two parts as one object, and with that also to keep unfitting combinations, which don't satisfy the perceptual constraints, out of the class of cups.

It may be necessary to distinguish two kinds of unitisation-type learning cases: associative chunking and perceptual unitisation.[4] Associative chunking is the process through which two elements that co-occur regularly become associated. If one is accustomed to getting a glass of water with an ordered cup of coffee, then being served just a cup of coffee will create the expectation of a glass of water that has to follow: DRINKING COFFEE is an activity-concept with these two elements. This is a straightforward example for associative chunking. Let me consider a case in which two feature dimensions are reliably correlated so that the occurrence of either one is a reliable sign of the occurrence of the higher-order phenomenon. Would such a case be more aptly described as unitisation or as associative chunking?

Suppose that fire fighters always take big, red vehicles that sound a siren alternating between the first and the fourth tone of a scale (say C and F), and that all other emergency sirens use a different interval, say the prime and the fifth (C and G). Upon learning about the visual properties of fire engines, one might form a concept FIRE ENGINE that is related to big, red vehicles. Having also learnt that the peculiar siren sound of the prime and the fourth is the fire fighter siren (having formed FIRE FIGHTER SIREN), one has formed the basis for putting together those two stimuli as the two most reliable signs for the presence of a fire engine. Thus, either of the stimuli can be used to trigger FIRE ENGINE, despite the lack of the other. While more elaborate than the coffee-and-water case, from Goldstone's perspective this would still count as chunking, since the co-occurrence is not based on spatial, but on causal and temporal contiguity, which supports the formation of two separate feature elements that later get a common 'heading' FIRE ENGINE.

The difference between unitisation and chunking can be made clearer by another example, this time adapted from Landy and Goldstone (2005, p. 352): in the same way that a photograph of a group of people is a combination of pictures of the individuals, a unitised concept is a combination of features that stand in certain cognitively interesting, complex relations – possibly with spatial configuration as the main combinatorial criterion (as in the photograph case). In light of the perceptual constraints on unitisation, which are the main point of difference to associative chunking, it would be prudent to not expect unitisation over different sense modalities such as visual and sound perception (as in the fire engine example) – at least until robust experimental data supports this idea.

4.3.2 Differentiation

The second mechanism for concept learning is dimension differentiation, "by which dimensions that are originally psychologically fused together become separated and isolated." (Goldstone, 2003, p. 249) Especially in differentiating dimensions, perceptual constraints influence the process: while it is easy for adult perceivers to separate

[4] This distinction was pointed out to me by a reviewer of Stöckle-Schobel (2012).

the properties size and brightness, it is much more difficult for the non-specialist to separate other fused dimensions such as brightness and hue. Differentiation might also be at work in separating non-dimension features, as in the fire engine example above. Suppose one never has paid much attention to siren sounds, and so has never noticed the difference between the fire fighters' siren and all other sirens – one has one single concept EMERGENCY SIREN. Upon learning about the tonal difference between the two intervals, probably in a music class, one might start noticing the difference, and thus differentiate one's concept into FOURTH INTERVAL SIREN and FIFTH INTERVAL SIREN, and then even relate these to the appropriate kinds of emergencies ("there's been a robbery next door, I'm quite sure that I'll soon hear the fifth interval police siren").

To make the differences between the perceptual-learning perspective and classical computationalism clearer, here is another rephrasing with an example. For the fixed-feature approach, new mental representations are new combinations of previously available primitive elements. Associative chunking, as in GLASS OF WATER and CUP OF COFFEE as components of the concept THINGS THAT I DRINK WHEN HAVING COFFEE, requires the availability of the components that are combined. In the variable-feature approach, new representations need not necessarily be primitive elements, or psychologically pre-available elements. Rather, they can be stimulus elements with "no parsing in terms of psychological primitives" (Landy and Goldstone, 2005, p. 350) – so, what Landy and Goldstone want to argue for is the import of perceptual tokens into the cognitive system. As with LEGO blocks, constructing the concept CUP from perception is like designing a LEGO cup, with the restrictions that this brings to the use of the concept; you can mainly use the cup for typical cup-involving activities like drinking and measuring, and not for activities like building a LEGO house from LEGO cups).

One already-alluded-to example comes from Burns and Shepp (1988)'s study on colour vision. Their main idea is that the three defining features of any given colour – its brightness (value), saturation (chroma), and hue – are difficult to selectively attend to for an untrained observer since colour perception is the perception of quite holistic stimuli. Their experiments demonstrate that test subjects have difficulty separating these dimensions when comparing a range of samples. In their study, Burns and Shepp also found that differentiating brightness and saturation is easier for trained individuals, such as artists. Landy and Goldstone take this as evidence for the creation of new feature detectors: if there only was one detector for COLOUR before the training, and the subjects were able to differentiate the brightness and saturation of a range of colours after their training, then the perceptual task must have been the cause for feature learning, and for the creation of new perceptual, discriminatory capacities. And surely, if a person didn't know the difference between the brightness and the saturation of a colour before, and could make a discriminatory judgment after the study, then a new concept has been learnt.

A final, important aspect of Goldstone's proposal is that his and his colleagues' studies don't rely on predetermined, fixed stimuli sets, but on totally novel ones that often cannot readily be parsed into already-known structural elements. An example can be found in Schyns and Murphy (1994)'s 'Martian rocks' studies. The study employed various black blobs with several kinds of round or prolate appendices, without any indication of possible fragments, or parts of the whole object. In their argument for using such alternative materials, Schyns et al. (1998) express that they want to exclude the possibility of using known categories in their experimental tasks, and that they want to get a better understanding of the ways in which totally new categories are learnt. If a shape is almost certain to not represent any possibly innate, fundamental shape primitive, then it should be very likely that learning to pick out that shape is a case of learning something new. With alternative materials, many different interpretations are possible; there are multiple features that could be encoded, and the analogue format (as opposed to the digital signs one also finds in fixed-feature experiments) makes it possible to study something akin to real-life concept learning, where the interesting or learn-worthy features also aren't as plainly recognisable.

With this picture of Landy and Goldstone (2005) reply to Fodor's challenge in mind, let me now confront the question if their proposal can stand up to the challenge.

4.4 Concept learning or conceptual change?

Given that the Perceptual Learning approach can do the things described above, does it answer Fodor's challenge for concept learning? I want to argue that it doesn't, because of two problems. Goldstone and colleagues have thus far left decisive questions pertaining to the central elements of their account unanswered, namely the details of the integration of perceptual symbols into the representational system, and the role of features and stimuli in that process.

4.4.1 Do unitisation and differentiation operate on concepts?

First and foremost, it isn't clear whether the model provides prospects for concept learning at all. One might agree that the phenomena of unitisation and differentiation are a form of learning, since they are mechanisms of restructuring previously available categories, and thereby they are means of grouping information in new ways that might lead to new beliefs. Consider somebody who finds out that two animals which she knew to be dogs belonged to two different breeds, say Labrador Retrievers and Dalmatians. This could clearly count as learning something one hadn't previously understood. But does it really count as introducing two new 'psychological primitives'? An alternative view along fixed-feature lines would be to grant that LABRADOR RETRIEVER and DALMATIAN are indeed new, but only as names for two objects that had already been processed in thought in a different way, say as p and q (the letters standing for token representatives of the individual dogs). So, what has been added were not new symbols, but rather new labels for old symbols, or new beliefs about these symbols, as in "p is a token of LABRADOR RETRIEVER".

To be fair to the Perceptual Learning approach, I propose to look at a more perceptually taxing kind of differentiation process, since this might support Landy and Goldstone's position. Suppose that, in a psychological experiment, a subject is rewarded for identifying tokens of pacman shapes with a 92° 'mouth' angle, and not rewarded if he chooses pacman shapes with a 90° 'mouth' angle. Wouldn't one want to say that learning to appropriately keep those two shapes apart counts as learning a new concept? While the example is intriguing, and representative of a class of psychological experiments on categorisation, I would argue that it doesn't count as a case of learning a new concept of, say, 92° PACMAN. Since the goal of the subject lies in getting the reward, it seems more appropriate to speak of the concept CHOICE THAT GETS ME A REWARD and the perceptual input that is related to a token of the concept – a choice of a given answer, say 'A' or 'B' (if 'A' and 'B' stand for the answers related to the respective 92/90° pacman). If one carries this thought further, the discriminatory input does not become involved in the content of the concept that is applied in the task – a 92° pacman is not a token of CHOICE THAT GETS ME THE REWARD, but it is a prompt to apply the concept by acting in a certain way (like saying "This is the right kind of shape"); at best, the pacman figure is reliably correlated with the ...REWARD concept. This is not to say that fine perceptual discriminations can never become conceptually relevant, or the topic of conceptual development: examples like Smith and Kemler (1978)'s study of changes in the integrality of dimensions such as colour and shape surely count as evidence to the contrary. The theoretical status of these developmental changes, however, is exactly the topic of this chapter.

Returning to the dog example, the tricky question for the fixed-feature computationalist at this point is not whether the new dog breeds were learnt, but rather how the symbol, the name, and the object that the symbol and the name denote are causally related. This question is related to Fodor (1998)'s discussion of the doorknob/DOORKNOB problem, and has been answered by appeal to a metaphysical and a neurological account of concept acquisition. In Chapter 2, I have raised problems for these answers, which an appeal to a computational fixed-feature language cannot fix. Perceptual learning can evade this metaphysical question ("How can an innate symbol refer to anything that was encountered perceptually?"), but at the price of creating a psychological one ("How is it possible to import new cognitive symbols from perceptual origins into a Language$_C$?"), which will create the second worry that is analysed further below.

4.4.2 Intermission: Kinds of learning

At this point, one might be tempted to postulate that there are several kinds of learning. In one kind of learning, some genuinely new primitive psychological token – be it a new feature, or symbol, or whatever kind of enrichment one might be interested in – is incorporated into the cognitive system. A case in point would be turning a perceptual stimulus into a cognitively usable symbol that can be used for category judgments, forming thoughts, or other conceptual tasks. Here, something that hasn't been part of the Language$_C$ would be transferred into that same language. Fodor's challenge is concerned with this kind of learning.

In another kind of learning experience, the available pieces of information get reordered, linked to other bits of information, or get categorised in a finer raster. Strictly speaking, nothing new enters the store of cognitive symbols, but the differentiation between different kinds of already available symbols will be finer, or coarser, depending on the type of change. The question remains if Landy and Goldstone would be happy with 'only' providing a model for the second type of learning, since the aim of their article, in their own words, clearly was to give a model for the first type:

> (...) our alternative to fixed-primitive languages involves not giving up computationalism, but enriching it with mechanisms which allow the construction of new psychological primitives that are not just combinations of other known categories. (Landy and Goldstone, 2005, p. 347)

On one interpretation of the Perceptual Learning approach, the main processes of unitisation and differentiation seem to fail to introduce new concepts, since they only operate on existing concepts, which are modified to be either more general or more specific regarding certain features of a given category. As Landy and Goldstone (2005) openly state, "feature creation simply involves alterations to the organization of stimulus elements into features." (Landy and Goldstone, 2005, p. 349) But a more strict computationalist, like Fodor (2008), could easily argue that this process does not strictly speaking add any new information to the cognitive system. Rearranging old concepts, on this view, cannot be counted as learning, since there is no new information added, but only a regrouping of old concepts. Like in the dog case above, there would only be the addition of new labels for objects that have previously been parts of the Language$_C$. If, however, one objects to this analysis and maintains that unitisation and differentiation mainly work on percepts, then the worries raised in Section 4.5 below will apply.

A variant of Fodor's hypothesis-testing paradox can be formulated that transfers the point into the feature-based learning Goldstone endorses. In order to categorise a stimulus as being evidence for, or being a token of, a certain psychological feature, one needs to know what feature that is – in order to perceive a sound as a fire-fighter siren, one needs to know what a fire-fighter siren sounds like (e.g. a fourth interval). And to know that, one surely needs a feature category that is available before having a stimulus to categorise accordingly. If there isn't more to what Perceptual Learning can do for our understanding of what concept learning should be, then it doesn't have an explanatory advantage to Fodor's Radical Concept Nativism in this regard, and needn't even be incompatible with his large and fixed basic vocabulary of the mind – the primitive symbols being given, while more elaborate concepts might well be composed by mechanisms like unitisation and differentiation. For example, a nativist might explain cases like Schyns and Murphy (1994)'s Martian-rock experiments not by stipulating changes in categorisation that were brought about by perceptual adaptation. Rather, she might link the categorisation success to activations of certain more primitive shape-concepts. Suppose that several detectors are in place to detect a limited range of primitive shapes. If a detector registers an occurrence of 'its shape', then the concept for this shape is activated. Now, one might say that in Martian rock displays, a range of these detectors signals the presence of this variety of different shapes, so that the concept that is linked to the detection of a Martian rock is the conjunction of all the primitive shape concepts. Improvements in recognition tasks will then not be traced to conceptual changes, but rather to changes in patterns of concept use, or something similar.

4.4.3 Features as concepts – a related issue

Gauker (1998) addresses a similar worry concerning Schyns et al. (1998). Gauker in fact poses a double dilemma to any 'concepts as (composed out of) features' approach. Suppose that concepts are composed of features, and that learning a concept involves learning a certain amount of the properties that are associated with the concept. Learning the concept BIRD might be linked to associating the concepts HAS WINGS, HAS A BEAK, FLYING ANIMAL, or any other combination of attributes, to the concept BIRD. Yet, this would require these attributes to be developmentally more basic than the concept BIRD. How could that be? Gauker sees only two possibilities, which pose a dilemma for Goldstone. Either one accepts that there is a developmental hierarchy of features and concepts. This option is in principle open to anybody, but has special advantages for fixed-feature theorists. One could postulate that there are certain primitive features at the basis of the more elaborate conceptual constructions that are learnt in cases like learning BIRD. These features are already parts of the cognitive system that didn't need to be learnt, and so would form an innate basis for our more superordinate concepts. If these features were innate, or pre-specified, they would be fixed. As laid out above, Goldstone wants to have some room for flexible features, so only relying on this option doesn't seem to be viable, especially since fixed-feature theorists might postulate a big enough, or flexible enough, primitive basis of concepts that could ground any supposedly perceptually learnable concept. The other option is to deny that feature concepts need to be more primitive than the superordinate concepts – Gauker associates Schyns, Goldstone and Thibaut with this view. The problem for this view is that it requires an explanation of exactly how "truly new features are created" (Gauker, 1998, p. 27) – features that don't have a previous history of, e.g., having been fused with other features, forming a less differentiated category.

4.4.4 Landy and Goldstone's reply: the Pask device

Landy and Goldstone attempt to answer these kinds of criticism by pointing out the changes that they have observed in their studies and in similar studies. They cite evidence that "early perceptual devices can be systematically and physically altered by the environment to change their representational capacities" (Landy and Goldstone, 2005, p. 351) to support the claim that new features can be created. For example, in simulations by Rumelhart and Zipser (1985), connectionist systems were able to create new detectors for different kinds of stimuli in a competitive learning task. But while the evidence might support this claim, it certainly needn't support the connected claim that such a change in representational capacities causes changes in the Language$_C$, and by this causes the learning of new features. A fixed-feature theorist might be happy with the first claim, linking it to the activation or triggering of a certain store of symbols that affects early perceptual devices: environmental influences would first cause changes in the (already-present) symbolic system, which would in turn result in changes in perception. The change in representational capacities thus might just be a change in frequencies of triggering certain symbols. One might call this a form of learning, since there would be changes in the perceptual domain, but the corresponding changes in the conceptual domain – starting to use previously available symbols for hitherto unperformed perceptual tasks – wouldn't be substantial enough to warrant the label 'concept learning'. Again, speaking of conceptual change – that means, speaking of changes to already existing concepts that don't amount to concept learning – would be more appropriate given the constraints of Fodor's argument.

Adding to this point, one can enlist another example of a learning system that Landy and Goldstone briefly introduce in their 2005 paper, and which they revisit in Goldstone and Landy (2010): the Pask device. A Pask device is "an array of electrodes partially immersed in an aqueous solution of metallic salts" (Landy and Goldstone, 2005, p. 351) that will physiologically change when electric currents are applied. Now, changes in electrical configuration in the device come with changes in functionality – the device will start reacting discriminatively to two kinds of sound frequencies: a 'new ear' for the circuit has been trained while it got constructed. From Fodor's perspective, it would however be a mistake to call this learning. He would argue that these changes have many of the characteristics that brute-causal acquisition of concepts in humans has. Most notably, that they are not mediated by beliefs and desires. So, by definition, they do not amount to concept learning. This issue is independent of the question whether the Pask device actually is a representational system – whether a certain reaction to frequency A counts as representing that frequency. The debate about the question "What counts as a representational system?" is a big topic in the philosophy of the cognitive sciences, and I have to restrict my engagement with it here. Prinz and Barsalou (2000) give a convincing argument that even some paradigm examples of non-representational systems, such as the Watt governor, should actually be regarded as representational systems if we concede that “representation involves information” (Prinz and Barsalou, 2000, p. 55). I will follow Prinz and Barsalou (2000) and will regard the Pask device as a representational system, and in that sense a fitting analogy to a cognitive system. To offer a preliminary evaluation of the merits of the Pask device, I want to raise the question whether the Pask device’s development of an electronic ear is more like knowing the difference between smoke and steam after being hit on the head, or more like learning from observation about the difference. For now, I submit that it is more like the former, and thus not a case of learning in Fodor’s sense. In Section 4.6 of this chapter, I will discuss whether we should put the Pask device’s type of learning on a scale between these two extremes.

4.4.5 Intermediate conclusion

Up to this point, the Perceptual Learning approach hasn’t succeeded in answering Fodor’s Challenge, since the alternative fixed-feature theory has been shown to give equally powerful explanations of phenomena like changes in representational capacity, while not having the problem of having to explain how new cognitive symbols could be created from perceptual materials. Also, the perceptual learning approach hasn’t given a full model for the latter task, and thereby only stands on a partial base of providing good explanation for the influence of the conceptual on the perceptual. There is however another set of conceptual problems that call for a resolution before the Perceptual Learning approach can get off the ground and before we can assess whether an inference to the best explanation would support the fixed-feature approach or the Perceptual Learning approach.

4.5 The notions ‘feature’ and ‘stimulus’

4.5.1 The Compatibility criterion

A second worry that directly follows from the first one has to do with the notion of features and stimuli in concept learning. In Goldstone’s theory, concepts are (created out of) features; features are created from stimuli. Stimuli are in a format that is supposedly compatible with the symbolic vocabulary of cognition. I submit that compatibility is a decisive criterion in this context, and I want to propose a formulation for a Compatibility criterion for Perceptual Learning.

Compatibility $=_{(def.)}$ A set of symbols S is compatible with a cognitive mechanism M iff inputting S into M yields a (symbolic) output S_O which can be used by further cognitive mechanisms. A set of symbols S is compatible to a second set of symbols S_2 iff S and S_2 can both serve as input for M individually, and iff symbols from S and S_2 in combination also yield a symbolic output S_O which can be used by further cognitive mechanisms.

Without fulfilling the Compatibility criterion, it wouldn't be possible to incorporate the perceptually based symbols into the previously available-and-exercised cognitive activities. Speaking in terms of Language$_C$, all symbols need to be combinable to form correct sentences from them.[5] Landy and Goldstone (2005) seem to concur with the need for something like a compatibility criterion when they emphasise the need for a language of stimuli from which to construct new features:

> (...) the claim that novel perceptual features can be learned sounds murky, or even mystical, without the clarification that the novel features are always drawn from a larger, more expressive, more primitive language embodying the physical and pre-conceptual constraints on what can be incorporated into features in the first place. (Landy and Goldstone, 2005, p. 348)

4.5.2 The proximity of stimuli

What is a stimulus then? In any dictionary of behavioural and cognitive science, one will find descriptions of perception in terms of proximal and distal stimuli. The distal stimulus of a visual experience might be the tree one sees, whereas the proximal stimulus would be the light reflection arriving at the eyes. At which level do Landy and Goldstone (2005) individuate stimuli? This question is pertinent to our present investigation since Landy and Goldstone should address it to make clear where exactly they would see the origins of perceptual experiences, and with that the origins of perceptually based concepts: are they in the world (i.e. distal stimuli) or are they in sensory activations (i.e. proximal stimuli)?

In Landy and Goldstone (2005), they discuss examples of roughly half-moon shaped figures combined of 5 segments and call these objects stimuli. In another example, they talk about "pieces of physical information [that are related to, or] packaged together in the same psychological feature" (Landy and Goldstone, 2005, p. 349) and use this as a synonym for 'feature'. These are exemplary descriptions of what Landy and Goldstone treat as features or stimuli, and these are the most concrete characterisations that they give of those terms. A look at earlier renderings of the theory might help. Schyns et al. (1998) offer the following characterisation of the meaning of the term 'feature':

> [The term] *feature* will refer to any elementary property of a distal stimulus that is an element of cognition, an atom of psychological processing. This does not imply that people are consciously aware of these properties. Instead, features are identified by their functional role in cognition; for example, they allow new categorizations and perceptions to occur. (Schyns et al., 1998, p. 1)

Here, first, features are described as 'elementary properties' of distal stimuli. They are also implied to be 'elements of cognition', i.e. Schyns and colleagues postulate a

[5] This point is related, though not identical, to the issue of the compositionality of thought, raised e.g. by Fodor and Pylyshyn (1988).

transition of perceptual properties into cognitive functions. The second point concerns a property's role in cognition – it is supposed to be functional. By this description, Schyns and colleagues want to counter the objection that some feature of an object might not enter a perceiver's conscious awareness and thus shouldn't count as an element of cognition. While the wording in the quote above suggests a definition in terms of distal stimuli, it also alludes to the psychological role of features, which is even more obvious in Goldstone (2003):

> A psychological feature (...) is a set of stimulus elements that are responded to together, as an integrated unit. That is, a feature is a package of stimulus elements that (...) reflects the subjective organization of the whole stimulus into components. (Goldstone, 2003, p. 242)

So, I suggest we should understand Landy and Goldstone (2005) as taking proximal stimuli as the background language. Still, there is the unanswered question how these perceptual signals are transferred into language-like symbols that can be used in the same cognitive operations as either innate or previously acquired symbols. How are these vocabularies matched to each other? Let me push the language analogy a little further with an example. Suppose the computational Language$_C$ is like English – approximately every English word corresponds to a mental symbol. Now suppose that the cognitive vocabulary gets enriched with a number of specialised features developed from a perceptual task, like learning chess moves. Perceptually learning a chess move, as opposed to learning it from a written description, might work as follows in the case of the rook's permissible movements on the board: straight lines along the horizontal and the vertical axis, but not on diagonals. The correct movements can be observed by watching rooks in a large sample of chess moves, or video clips from chess matches, also with a variety of token rooks (made from different materials, or shaped in a variety of ways), and one might even learn how to tell whether a chess piece is a rook or a queen. The interesting question then is: would a thought about a situation in a chess match be multimodal – would it involve the perceptually learnt symbol for the rook as well as the previously available non-domain-specific vocabulary? Let's take "[]" as a replacement for the perceptual symbol related to the rook moving one square to the left, just for this example, and phrase a thought like "If the rook moves one square to the left, the players will stop playing" multimodally: "If [], the players will stop playing" (given e.g. that the result is a checkmate). Is it possible to infer the consequent of this conditional from being presented with a representation of []? While this last question might be for further empirical studies to decide, it already hints at the more general worry about perceptual tokens of some sort and their role in cognitive operations: given that originally, a certain cognitive function is performed by a mechanism using symbols of a (possibly innate) Language$_C$, how can the mechanism adapt to new symbols being introduced into it and filling that cognitive function? That is, how can a perceptual symbol store and an innately fixed symbol store become *compatible*, as defined above? Landy and Goldstone don't offer a model for this, and so I conclude that, as it stands, the construction of variable-feature language hasn't been sufficiently based on a model of transferring perceptual symbols into conceptual systems.

4.5.3 Feature detectors and maturation

Sticking to the notion of features as primarily relevant to building mental representations, one could bring the theory of feature detectors into play, as in Barlow (2001). This, specifically in visual perception, would be an obvious way out of the problem. Yet, this raises another problem – the feature detectors would have to be tuned to some specific inputs. But how did the feature detectors come into being, and how did they get the tuning they exhibit?

Feature detectors are an instance of (possibly innate) processes that one might use to explain concept learning, under the assumption that concepts needn't be built into a system as long as there are built in processes that will be able to import concepts into the system. The problem, however, remains the same: a nativist can always argue that built-in processes need to be tuned to their inputs in some way. In the Perceptual Learning literature (broadly defined), one can find several references to Gestalt laws (Schyns and Murphy, 1994; Bhatt and Quinn, 2011; Quinn et al., 2006), and specific proposals that acquiring Gestalt laws is possible through neural network simulations (Gerganov et al., 2007). Yet, how should adherence to a given Gestalt law, e.g. good continuity, be possible for a cognitive system without having concepts that are able to express the law and that would help classifying perceptions according to the principle?

Another proposal would be to appeal to maturation as a factor in Perceptual Learning. But maturation, or another form of innateness, would yield no new variable-feature language in the sense of Landy and Goldstone, as its characteristic feature is its independence of the specific type of stimuli the system is confronted with. A classic example of this could be imprinting in newborn ducks, as discussed by Fodor (1981): any moving object will trigger the 'concept' MOTHER, having the duck following the moving object. The important thing is that the duck's sensory system is predetermined to follow the closest moving object.

If on the other hand the specific stimuli play a role, or matter in some sense for the creation of psychological primitives, how do they influence the creation of a feature detector? This is a main question we should be asking when looking at concept learning, and for which we don't seem to have found an answer in the Perceptual Learning literature reviewed in this chapter yet.

4.5.4 Physical information

We seem to have gone back to the original question, and the appeal to a larger background language of stimuli hasn't advanced us very far. So, instead of stimuli, one could try and look at the perceptual part of concept learning more abstractly, as the transfer of information. After all, Landy and Goldstone also refer to the materials from which to get new features as “physical information” (Landy and Goldstone, 2005, p. 349). If we could make Landy and Goldstone’s point without directly referring to stimuli, then we might avoid the questions that we raised above.

Now, the first task is to disambiguate the notion ‘physical information’. Either, physical information, or a bundle of features, is supposed to be informationally structured before or in being perceived. This would require a form of direct realism, or of direct perception: for example, a given object affords to be perceived as a tree, so we as perceivers pick up the right kind of information in order to treat it as a tree. Or, on the other hand, physical information is the cognitive content that has been extracted from the experience; this could be something like a representation of a tree. The retina has registered a certain image, sent it to the visual cortex in one way or another and there, a representation of the tree is formed, or accessed, or activated. In line with the above decision to talk of stimuli in terms of proximal stimuli, it seems sensible to choose the interpretation of physical information as cognitive content. This interpretation, however, just skips the interesting question, which is “How does a representation of a new object come to be included in a cognitive system?”, and leaves the field open for a Radical Concept Nativist reply to the effect that the representation was triggered by some process, or that the perceptual stimulus got paired with an arbitrary symbol from the wealth of symbols in the representational mind. To avoid this, any proponent of Perceptual Learning has to show just how perceptual content enters cognition and by which means new symbols, or new bits of a feature-‘language’, are added to the system. Landy and Goldstone don’t offer a model for this form of feature learning, and so I have to conclude that – while it isn’t conceptually excluded that such a model is possible – their proposal still has some way to go before it can pose a fully developed alternative to their fixed-feature opponents.

4.6 Perspectives for an amendment of the Perceptual Learning approach

4.6.1 Three kinds of percept-concept interactions in Perceptual Learning

As discussed above (in Section 4.4), the main problem in the Perceptual Learning account is that Landy and Goldstone don’t seem to make their project’s scope sufficiently clear: while they express the desire to explain the formation of new concepts from perceptual materials, they seem to rely on conceptually available tokens (processed stimuli?), or even primitive concepts to do the formation work. On this interpretation, the role of perception in concept learning isn’t clear, and a Fodorian might reasonably argue that it doesn’t play any role in it. To cast away this difficulty, and assuming that there is a feasible way of explaining the learning of new concepts within Landy and Goldstone’s framework, let me sketch three ways one could understand the relation between concepts and perception in this theory.

The first case would be that perception is the input to the learning processes that turn percepts into concepts: after certain stimuli have been united or differentiated, they have come to be concepts. Landy and Goldstone seem to aim for a system that can achieve this kind of learning. Above (in Section 4.4.), I have argued that this possibility hasn't been developed enough by Landy and Goldstone, since they don't offer an explanation of the way in which the transformation happens.

Second, concepts could be the input to the learning machinery. An existing concept, which might even be an innate concept, can come to be 'split' into several more precise, or more applicable concepts. I have argued that, on one interpretation, Landy and Goldstone seem to offer a new model for this, and have criticised this interpretation of their theory as an insufficient response to Fodor's Challenge, since Fodor could object that something that has been built from pre-existing concepts cannot have been learnt. As an added problem for an account of concept learning that wants to explain the role of perception with this kind of learning process, perception can easily be construed as playing a minor role in this process. Fodor's idea of perceptual triggering is one instance of possible perceptual influence here, and Fodor's examples like learning GREEN AND RECTANGULAR after seeing an object that is green and rectangular are ample demonstrations of the problem.

There is, however, a third option: perceptual and conceptual mechanisms could work together in a way that creates new concepts from available ones with guidance from perceptual tokens. This idea has to be spelled out carefully, and the differences with the other two types of concept-percept relation have to be highlighted.

First, for this third type of interaction, one has to assume that percepts don't become concepts to avoid the problems of the first type of learning. Even though we want to tell a story about how seeing a leopard can lead to learning the new concept LEOPARD, we will deny that the image itself will assume the role of a 'leopard concept'. Perceptual tokens, like an image, thus are not transformed on this account.

Second, one has to assume that percepts play an important role in concept learning – a role that goes beyond Fodor's idea of triggering. This is needed to avoid the problems of the second type of percept-concept interaction. What could this look like then?

Perceptual learning might create new concepts from old concepts by uniting and differentiating perceptual tokens that have been classified as instances, or exemplars, of previously available concepts. For this kind of interaction, it is necessary to assume the following. Concepts are used to classify (perceived) objects – we apply them in engaging with the way our perception represents the world to us. This is a fairly uncontroversial assumption. It can be reformulated by stating that perceptual tokens are subsumed under, or judged to be instances of, certain concepts. This reformulation invites the picture of concepts as labels that are stuck to files, and of percepts as part of the content of these files, with the possibility of other 'file contents' such as rules for the use of a concept, or beliefs about a given concept, or further files of subordinate concepts. Applying a concept in perception would for instance involve checking whether a perceptual representation should be grouped with the other perceptual representations that are already part of the concept's file. Applying a concept in an inference would involve checking for permissible moves in the set of inference rules stored in its file.

This would deserve the name Perceptual Learning much more than the second option would, since it operates on perceptual tokens, albeit on pre-categorised ones. An example of this can be made in the fire-fighter siren case: suppose the differentiation case of starting with a broad concept EMERGENCY SIREN, which doesn't discriminate between police, fire fighters, medics, or indeed any kind of emergency siren. Yet, one might have heard all three sirens quite often, and could have used this sensation to guide one's behaviour, e.g., by pulling over the car to let a police car pass. As in the example given above, on might learn about the musical difference between the two kinds of siren (first-fourth and first-fifth interval). Upon training one's senses to be more sensitive to such musical differences, one might come to the point of evaluating the different kinds of emergency sirens, and might judge that a finer classification will be appropriate. The fourth interval only being used by fire fighters, while all the other sirens use the fifth interval, one will reclassify one's auditory perceptions of sirens, and will have two categories emerging from this process – fifth-interval signals and fourth-interval signals, with FIRE FIGHTER SIREN being a sub-category of the latter.

Does this interpretation fare any better against Fodor's Challenge? In one sense, it still doesn't escape the paradox, because it makes concept learning depend on available concepts. The effect of learning a newly unitised or differentiated concept could be explained as analogous to the GREEN AND RECTANGULAR case.

In another sense, however, this third type of percept-concept interaction escapes the paradox because it is able to explain how perceptual changes can lead to changes in conceptualisation – to changes in judgements, or to the attribution of category membership. It can explain how the combination or dissociation of two concepts can be informative rather than trivial. In informative differentiation cases, a token of perceptual information related to a concept might not be completely compatible to the concept, and prompt a division of the concept's instances that is based on the perceivable differences that first became apparent when that token was perceived and categorised. An informative case of dissociation comes from the emergency siren case, where further musical education brought about a differentiation between two kinds of siren sounds, whereas a trivial one would be getting GREEN and RECTANGULAR from GREEN AND RECTANGULAR.

Before concluding the treatment of Goldstone's Perceptual Learning approach, I want to look at another response he can give to overcome the concept-learning challenge.

4.6.2 Blurring a boundary: brute-causal and rational acquisition

When looking for amendments of the PL approach to make it overcome Fodor's Challenge, taking up the idea of the Pask device as discussed earlier might be instructive:

should perceptual learning in humans count as rational learning or as brute-causal acquisition? With the verdict from above concerning the Pask device's being more like being hit on the head than like learning, one might want to opt for the brute-causal option. While this may sound like a step back, or even like waving the white flag, it can in fact turn out to be a saving move for Perceptual Learning's role in concept learning. This point can be developed into an argument against the methodological set-up of Fodor's Challenge – the distinction between brute-causal and rational acquisition might not be as helpful as Fodor would like it to be, and cases like the Pask device might work in favour of giving up the distinction altogether. To make this clearer, let's have a look at what Fodor takes to be non-rational acquisition in LOT2. He gives a list of things that might count as acquisition, but not as learning:

1. "*There are all sorts of mind/world interactions that can alter a conceptual repertoire* (...): sensory experience, motor feedback, diet, instruction, first-language acquisition, being hit on the head by a brick, contracting senile dementia, arriving at puberty, moving to California, learning physics, learning Sanskrit, and so forth indefinitely" (Fodor, 2008, p. 132)

2. Classical cases of non-rational acquisition, for Fodor, are "surgical implantation; or (...) swallowing a pill; or (...) hitting one's head against a hard surface, etc." (Fodor, 2008, p. 135)

3. Fodor is very clear about one decisive consequence of his argument: "(...) acquiring a concept from experience must be distinguished from learning it." (Fodor, 2008, p. 145) As discussed in Chapter 2, he goes on to develop a 'neurological' theory of concept acquisition (Fodor, 2008, pp. 146-168) that uses prototype learning as the source of experientially grounded concepts.

Fodor wants to show that concept learning isn't needed for the acquisition of concepts – in his view, it just isn't the right thing to look at when we want to get an understanding of why we have the concepts that we have. The quite disparate list of acquisition processes in (1) stands as a kind of testimony for this. It raises the question: is there anything else, besides not being learning for Fodor, that unites these instances of acquisition? Or are there reasons to reject (1) as an appropriate and informative grouping of mechanisms and events that could explain the phenomenon of concept acquisition? Items like sensory experience, motor feedback, instruction, and first-language acquisition have a central role in human development. In several theories, these types of phenomena are used as instances of concept learning (see Chapters 3, 5, and 6 for examples of such theories).

In comparison, moving to California, or learning Sanskrit, isn't necessarily part of an average person's development (although they can be seen as tokens of more generally spread behaviours, like 'moving to a new place' or 'learning a second language'). Still others, like one's diet, or blunt trauma to the head, are quite a bit further removed from the realm of what can meaningfully be called 'conceptual activity'. I want to propose that there is a continuum of cases ranging from ‘completely non-rational’ to ‘Fodor-standard rational’, and that Fodor draws his line at the wrong point: he separates the latter from all other cases, including those that have a clear bearing on issues of concept learning as a rational enterprise. If there needs to be a line that is drawn between cases, then it would be sensible to draw it between ‘completely non-rational’, as in miraculous hit-on-the-head cases, and all the rest. I use the example of the Pask device to illustrate this point.

Even though one might liken the Pask device to the blow to the head, and to its possible effects on the conceptual system, there are clear differences between the two. First, the causal chain between the environmental influence on the system and the ‘cognitive effect’ of it is much more direct in the Pask device. Electrical currents influence the concentration of cations and anions in the vicinity of the electrically charged wires, and this in turn influences the formation of new metallic/electronic connections in the device. At a certain point of this process, the device has ‘learnt’ to react differentially to two kinds of sound wavelengths. The relation between a knock on the head and the acquisition of a given concept X is far less clear, in comparison. This first difference of the directness and observability of the causal relation between cause and (acquisition) effect can already be taken as decisive for the verdict that we are looking at a difference in kind in these acquisition cases. Let us assume, further, that the Pask device’s electrical changes that lead to the acquisition of an ‘ear’ are sufficiently similar to some aspects of human sensory development.

Then, I want to propose that learning a perceptual discrimination of the Pask-kind can be sufficient for acquiring a concept that is reliably related to the perceptual input. Furthermore, if the perceptually acquired concept also obtains a cognitive role, in the formation of discriminative judgements, or of other kinds of beliefs about the object or class that the concept represents, then I think it is fair to say that the concept has been learnt through a combined perceptual and cognitive process. If one assumes that the perceptual representation itself becomes a concept, then the question remains if this sort of transition-process can fulfil the Compatibility criterion. If, on the other hand, one follows the proposal above and assumes that there can be a third kind of percept-concept interaction, then we have a representative case of concept learning on the basis of Perceptual Learning. It goes beyond Goldstone’s Perceptual Learning approach in so far as it draws a stricter line between perception and conception while maintaining the importance of perception for the learning process. It does this by grouping perceptual tokens as contents of a concept’s ‘folder’, while still not identifying them with the concept – the ‘label – itself. In that sense, perception and conception are still separate, despite working together closely.

Also, it helps us to cache out the LEGO analogy: based on a learnt perceptual discrimination, we can apply a new, in the sense of ‘not in use yet’, symbol of the Language$_C$ as the representation of the perceptually determined class. Like with real LEGO, something has to be a ‘template’ for the special-purpose LEGO piece, e.g. a picture of the sails of a real pirate ship. It should however be noted that the analogy doesn't carry too far: while the LEGO sails *are* designed to look like real sails, the special purpose symbol doesn't need to *look* like the object it designates via being linked to the perceptual representation of the object. All that is needed is the capacity for perceptual discrimination, together with some kind of perceptual representation that has been stored while developing the capacity (we can leave the exact format of this representation open for now – it might be a prototype, or an exemplar, or something related), and an as-of-yet-unused cognitive symbol that gets its content from being linked to this perceptual representation.

In conclusion, the 'non-rational' Pask-type learning can be seen as having effects that are far more controlled and far more reliably linked to conceptual changes than Fodor's prime examples of brute-causal acquisition. Challenging Fodor's argument on the grounds of overthrowing the distinction 'non-rational/rational' can be a part of a counterargument that attempts to explain concept learning within a largely computational framework. I will revisit this issue again in Chapter 6, when I reconsider the idea of a continuum between brute-causal concept acquisition and full-fledged concept learning.

4.7 Conclusion

At the end of my discussion of Landy and Goldstone's PL approach, I take the following points to be the main results. Perceptual Learning as a concept-learning mechanism faces two major problems for which I haven't found solutions in Landy and Goldstone's work. I am however not denying that PL might successfully overcome Fodor's Challenge. To amend the theory, one would need to say more about the input systems to unitisation and differentiation, and be clearer on the representational format that they are able to operate upon. Specifically, the following questions are still unanswered. How can a cognitive mechanism that was presumably first stocked with innate computational symbols grow to work with learnt perceptual features as input to, and vocabulary for, its activity? And is it possible to mix symbols of different origins and formats (amodal/modal) – to have 'multi-lingually' integrated cognition?

Until these issues have been addressed, the proposal doesn't deploy its full potential to threaten a fixed-feature approach à la Fodor. Even if both approaches can be construed as having similar levels of explanatory power, one of them doesn't face Fodor's Challenge because it accepts the Radical Concept Nativism conclusion, while the other doesn't yet overturn Fodor's empirical premise. The disadvantages that stem from the problems identified in this investigation weaken the Perceptual Learning approach's appeal and thereby put its opponent in a stronger argumentative position for now. However, with some further clarifications and a change in argumentative focus to the issue of 'brute-causal' elements in concept learning, one can construct a position that is close to Goldstone's, and can be developed into a viable theory of perceptual concept learning.

After this discussion, one might also be tempted to conclude that the notion of a feature-language, flexible or variable, is misguided as it invariably brings the issue of translation into the debate. How does the translation of a stimulus (and *which* stimulus) into a mental symbol work? Also, it suggests an ordered, or 'grammatical' structure in the non-mental/physical world that is the object of perception. This is dangerous, because the world as appearing to us might not actually be best carved into the perceived (natural?) kinds, but into theoretical kinds that we only perceive through mediation. One doesn't see the chemical structure of object x without being somewhat of a trained chemist, if at all, or at least it is not clear in the Perceptual Learning approach whether the interaction between perception and cognition leads to such depths of theory-ladenness of perception (as in *seeing* chemical structure) as opposed to a quick-and-dirty inferential connection between perceiving certain visual properties of a chemical sample – e.g. observing a deep green flame when burning a sample of a chemical powder, and identifying it as copper(II)-sulphate (maybe a perceptual-cognitive process), as well as the sample's being a salt (an inference from that observation). In making the distinction between perceptual-cognitive and inferential, the adherence of a given process to perceptual constraints might indicate that the process is of the former type, whereas observations following theoretical, or 'conceptual', rules can be properly classified as the latter.[6] Still, the distinction is not always a clearly cut one. This doesn't just touch upon Landy and Goldstone's proposal, but more generally on their opponents: if the language metaphor doesn't work for features, as one might conclude from the problems raised in section 4, then why should one be inclined to see the strong analogy between computers – symbol crunchers – and human minds with brains and nervous systems underlying them (in one way or another) as necessary? Perhaps the mind only becomes symbolic by consciously starting to use symbols, but this doesn't reflect this symbol-mindedness in the unconscious elements of cognition. Dissociating materials, or vehicles of thought, on the one hand, and thought contents on the other hand, might be a prudent move until a clearer picture of the connections between vehicles and contents is available.

[6] This was pointed out to me by a reviewer of Stöckle-Schobel (2012).

Chapter 5

Sellarsian concept learning

5.1 Introduction

In the preceding chapters, we have investigated psychological theories that reply to Fodor's Challenge of concept learning, with the intermediary result that neither of the ones we discussed gives a complete solution to it. In this chapter, I introduce Wilfrid Sellars's theory of concept learning, and connect it with a compatible strand of contemporary developmental research. While the discussion of previous chapters has been centred on Fodor's empirical premise regarding the hypothesis-formation-and-testing (HF) model, another approach to Fodor's Challenge deserves a closer look as well. I propose to challenge the third premise of Fodor's paradox – which specifies a sufficient possession condition for concepts – in a model of concept learning. To recall, Fodor's model for this is that

> a sufficient condition for having the concept C is: being able to think about something *as (a)* C (being able to bring the property C before the mind as such, as I sometimes put it). (Fodor, 2008, p. 138)

If we follow the argumentative strategy of calling this condition into question, we arrive at a different kind of concept-learning mechanism that promises to capture an important dimension of concept learning, as I shall argue in this chapter.

5.2 The Augustinian Conception

Doubts about the availability of an innate stock of concepts have guided a large tradition of research, and have recently been expressed in similar ways by Wittgenstein (1953) and Sellars (1956). Both argue against the so-called 'Augustinian conception of learning' (henceforth *AC*), which Fodor explicitly endorses as part of his challenge for concept learning. Sellars and Wittgenstein react to different aspects of AC, and both offer alternative accounts of learning concepts in learning language. Fodor and Sellars both take up Wittgenstein's thoughts on Augustine, and discuss the topic of the innateness of concepts in more detail than did Wittgenstein. I want to focus this chapter on Wilfrid Sellars's reaction to AC. After an appraisal of Sellars's theory of concept learning and of some apparent problems with his take on prelinguistic infant cognition, I will suggest some improvements and will review some links to contemporary developmental research. This will culminate in proposing a developmental CLM that is consistent with Sellars's theoretical commitments and which is based on research in social cognition.

For historical accuracy with regards to the contents of AC, I want to start the chapter with a quote from Augustine's *Confessions*, which is at the origin of the present discussion.[1]

[1] Although Augustine has also written in more detail on the subject of learning, in *De Magistro*, I will omit a detailed discussion of this work for now, except for some short remarks.

> I ceased to be a baby unable to talk, and now was a boy with the power of speech. I can remember that time, and later on I realized how I had learnt to speak. It was not my elders who showed me the words by some set system of instruction, in the way that they taught me to read not long afterwards; but instead, I taught myself by using the intelligence which you, my God, gave to me. For when I tried to express my meaning by crying out and making various sounds and movements, so that my wishes should be obeyed, I found that I could not convey all that I meant or make myself understood by everyone whom I wished to understand me. So my memory prompted me. I noticed that people would name some object and then turn towards whatever it was that they had named. I watched them and understood that the sound they made when they wanted to indicate that particular thing was the name which they gave to it, and their actions clearly showed what they meant, for there is a kind of universal language, consisting of expressions of the face and eyes, gestures and tone of voice, which can show whether a person means to ask for something and get it, or refuse it and have nothing to do with it. So, by hearing words arranged in various phrases and constantly repeated, I gradually pieced together what they stood for, and when my tongue had mastered the pronunciation, I began to express my wishes by means of them. (Augustine of Hippo, 1961, I, viii, 13)

This quote, or selections from it, has been taken as the foundation for the discussion of Augustine's contribution to discussions regarding the nature of the human mind, and the way humans learn languages. So, for instance, in Fodor (1975), Fodor links his arguments for the existence of an innate language of thought to Augustine's theory of language learning, which Wittgenstein (1953) had famously criticised. Let's go through this carefully, since it sets the scene of the present investigation. In the passage above, Augustine reconstructs the process through which he thinks that he learned to speak. As an infant, he had a variety of thoughts and desires, which he wanted to communicate. Yet, he couldn't, since he didn't know how to use the adults' language. But he had a god-given *understanding* of what the words *meant*, so he only had to map the uttered words to the objects he attended to to 'learn a word', broadly construed. Now, since this understanding of the meaning of the words was already there before the connection to the sounds of those words was made, one can attribute the idea of an innate language to Augustine. After all, he must have been able to formulate his thoughts and desires in such a way that they would connect his attitudes to the objects of his attitudes like words of a sentence are connected as subjects, predicates and objects. At least this is what Fodor takes Augustine to mean. In sum, Fodor endorses two hypotheses when committing to AC:

AC (a) we understand meanings of spoken words through ostensive definition, and (b) first-language learning is like second-language learning (cf. Erneling, 1993, p. 349)

Both parts of this formulation are ambiguous: (a) is not supposed to mean that a learner only grasps a given meaning after she was directed towards the object that is related to that meaning – quite the contrary. The meaning is already mentally available for the learner. What has to happen in such a pointing episode is that the word gets coupled to the object – the word has to be established as the trigger for the mental token (meaning, or concept) that is already *known to be referring to* that object in the world. In a way, it is as though there already is a mapping between the world (and its objects and properties) and the human mind's contents: any given mental token m already refers to a given worldly thing w – this is the 'god-given' part Augustine refers to in the above quotation.[2] The only thing which is not yet in place is the connection between the word designating w and meaning m – the learning of this connection, Augustine claims, is not due to any teaching efforts by his caregivers, but to his attending to their ostensions and his mental effort of linking the sound to the meaning.

Equally, (b) is a shorthand formulation of the following view, which is entrenched in Fodor's computational theory of mind: to learn a spoken language is to learn the correct relation between one's language of thought and that spoken language. The language of thought – the concepts that one possesses – is innate; it is the first language everyone has available, even though it is not learned. Every other language, such as the first spoken language, needs to be coupled to that language of thought, and thus can be seen as only the 'second' language one has available.

The second part of AC is more visible in the discussion of Fodor (1975), but it is enmeshed with the first hypothesis: an act of 'learning' a new word by linking it to the correct symbol in the language of thought requires a referential relation between word, symbol and denoted object. Since the denoted object, as we have seen in our earlier discussion (see Chapters 2 and 4), merely plays the role of a trigger, the requirements for the learning episode aren't very high, and can just as well work with the idealisation that word learning predominantly happens in situations where teachers utter a word while, say, gesturing towards its referent for the learner – although, as we have seen already, neither Augustine nor Fodor would call this *learning* in the proper sense.

Fodor (1975) gives several arguments against Wittgenstein's views, and he is particularly intent on refuting Wittgenstein with regard to his criticism of AC. Here is Fodor's argument for the claim "that Augustine was precisely and demonstrably right and that seeing that he was is prerequisite to any serious attempts to understand how first languages are learned" (Fodor, 1975, p. 64):

> Learning a language involves learning what the predicates of the language mean. Learning what the predicates of a language mean involves learning a determination of the extension of these predicates. (Fodor, 1975, p. 63f.)

Now, from the Wittgensteinian perspective, learning the latter involves learning that a predicate P falls under certain rules R. But Fodor raises the point that one can't learn

[2] As Burnyeat (1987) emphasises, Augustine grounds his view that we never learn anything from anyone else on the belief that all learning happens within oneself; specifically, we 'learn' because god has given us a mind with which we can understand the world around us.

what P and R mean without a language in which P and R can be represented. Fodor takes this to support his hypothesis that there has to be a language of thought, the terms of which are used to learn a spoken, or public, language. This language has to be able, as we just saw, to express any predicate of the public language that is learnt, because otherwise the predicate could not be learnt at all. So, the language of thought needs to have the maximal expressive power, and cannot be enriched by the addition of new symbols. The second tenet of AC is thus implemented in Fodor's reply to the concept-learning challenge.

Fodor sees this as a direct refutation of Wittgenstein's alternative to AC because the argument shows that a model of word learning by following rules succumbs to the same paradox as discussed in Chapter 2. Fodor's argument ties in with a distinct property of his theory of mind, which we can describe as deeply entrenched in the Rationalist tradition that Fodor often associates himself with: the 'Rationalist' possession condition for concepts. According to this possession condition, possessing a concept of something is 'to be able to bring something before ones mind as such'. This is exactly what is going on in AC, where an understanding of a word's meaning is the precondition for making the correct word-meaning-signified connection. I will take this as the relevant and necessary background from which to investigate Sellars's criticism of the Augustinian Conception, and its effects on Fodor's position.

5.3 Sellars's argument against AC – a reconstruction

In Sellars (1956), we find a short paragraph in which Sellars attacks AC as an instance of the Myth of the Given, which is his overall target throughout the essay. While most attention regarding the Myth has been directed at the critique of Foundationalism as a theory of knowledge, there is a similar worry for the philosophy of mind, which shall concern us here. The driving question is: how can a person become aware of the categorial status of a given object in her experience? Or, in simpler terms, can I know that a thing in my visual field is an x without knowing it under the description 'x' (*as* an x)? In the Augustinian picture, this is not a problem, since all understanding of what kind of object x is is already given to us by having the kind of mind that humans have. The only thing that has to be added to this understanding is the name that people use when talking about x, and learning this can happen through imitating the sounds that others make in reference to x. From Sellars's perspective, this answer is highly unsatisfactory, since it presupposes

1. the availability of the categories proficient language users and even specialist scientists have at their disposal (leading to the seemingly absurd view that infants already know what a carburettor is before even seeing a car, and that people in Plato's time already had the concept 'quantum theory')[3] and

[3] See Churchland (1978).

2. the correctness of the reference in any reference-fixing event to be built into the theory; if a concept only gets triggered by good instances of it, then misapplication of a concept, or learning from bad examples seems to be ruled out.[4]

I take these two items to be the main problems for AC. Below, I will put them into more context by posing three challenges to AC. These challenges will flesh out why the two items above are undesirable elements in a theory of concept acquisition. I will first give a clear formulation of the Sellarsian position with regard to the Myth of the Given as exemplified in the Augustinian Conception, and then go through the Awareness challenge, the Holist challenge, and the Relational challenge.

5.3.1 The Awareness challenge

In Sellars (1981a), he reformulates his definition of what the Myth amounts to in a way that fits AC quite precisely:

> If a person is directly aware of an item which has categorial status C, then the person is aware of it as having categorial status C. (Sellars, 1981a, §44)
>
> *To reject the Myth of the Given is to reject the idea that the categorial structure of the world – if it has a categorial structure – imposes itself on the mind as a seal imposes an image on melted wax.* (Sellars, 1981a, §45)

The notion 'awareness' as used by Sellars would profit from some illumination, to avoid the kinds of misunderstanding that often come with common sense terms being used to do philosophical work. In the system of Sellars (1981b), one can be aware *of* a green thing without being aware of it *as* green. This means that the colour of the object can play a causal role in how one reacts to it without playing an explanatory role for oneself. For a rat in a behavioural experiment, it might not make a difference that the symbol on the door that it has to choose to get a reward is a triangle. Presumably, the rat sees the triangle, but not *as* a triangle. Still, as soon as the rat has learned to take the triangle-door, the triangular sign is playing a causal role in the experimenter's explanation of the rat's behaviour. Within AC, this distinction is not made, and any kind of *awareness of* is actually *awareness as*. So, from Sellars's perspective, AC is too permissive in its attribution of concepts: simple stimulus-response pairings in any kind of animal that is able to learn these pairings would yield concepts in these animals, as in the rat following the green triangle sign. We can thus see that AC, and Fodor's concept possession criterion, are in need of a justification as to why the distinction doesn't play a role in this framework. If the distinction is not made, then the second issue I raised above applies and Fodor wouldn't be able to explain misrepresentation, or acquiring a concept from an atypical instance, or from a non-instance. Consider the case of acquiring ZEBRA from seeing a white horse with painted-on black stripes in a

[4] This point is an instance of the worry about the possibility of misrepresentation in causal theories of reference, see e.g. Millikan (1991); Dretske (1993).

situation that usually features genuine zebras, not fake zebras. The result of such an acquisition episode (independently of whether we will regard it as learning) should be the acquisition of ZEBRA, unless we change the context of the episode. If the distinction between 'awareness of' and 'awareness as' is made, then the misrepresentation that is involved in the process can be taken into account in one's explanation: the subject was aware of a painted white horse, but wasn't aware of it *as* a painted white horse, but as a zebra. This explains why the acquired concept will be ZEBRA rather than PAINTED WHITE HORSE.

Without the distinction, in cases where all awareness of a thing is awareness of it as that thing, we wouldn't be able to explain the error. We could hold that the person is aware of a zebra, and therefore acquires ZEBRA, but this would not solve the problem because it simply ignores that there was an error in the acquisition process. Alternatively, we could hold that the person is aware of a zebra, but acquires PAINTED WHITE HORSE because the acquisition process was triggered by a painted white horse. This is not a plausible answer because we should expect a correspondence between a state of awareness and an acquired concept. A similar point holds for the proposal that the person is aware of a zebra, but acquires ZEBRA OR PAINTED WHITE HORSE. I take it that this is a serious explanatory lacuna in AC, which at least weakens the appeal of AC as an adequate account of concept acquisition. The problem is similar to one of Prinz's objections against Fodor's metaphysical solution to the d/D problem, which I introduced in Chapter 2. With this context in mind, I propose to call the first element of Sellars's critique the Awareness challenge:

Awareness Explain why the distinction between 'awareness of' and 'awareness as' is not needed for AC, and how AC can account for error and misrepresentation in concept acquisition without it.

5.3.2 The Holist challenge

Another element of Sellars's critique can be called the Holist challenge to AC. A full-fledged concept would be part of a network of propositions, as can be shown by the colour example: when somebody masters the concept GREEN, then they are in a position to judge of any (unicoloured) green object that it is green, and additionally, that it is not blue, not red, and so on. I take it that a complete account of concept possession should include a condition that specifies the role of the inferences one can make when ascribing a concept. Sellars would for instance demand that one is able to make at least some inferences from the ascription of a given concept. If one has the concept ZEBRA, one should minimally be able to classify zebras in a different way than non-zebras in some circumstances; else, the concept wouldn't have a role in one's cognitive system. The point of this is that, even for a Conceptual Atomist like Fodor, who claims that "most lexical concepts have no internal structure" (Fodor, 1998, p. 121), concepts are used within a context of thoughts or of propositional attitudes. This implies that there are standards of correctness for the use of concepts; in a minimal sense, this means that one should be able to identify a concept by a given set of inferences one can draw from applying it to a given situation. This is already established if one can apply NOT GREEN to all things that aren't green.

Even if one wants to maintain that the ability to make inferences only has to be rather limited, as for instance in the examples given above,[5] then the Fodorian/ Augustinian position immediately becomes highly implausible as it would require an immensely rich innate conceptual repertoire. It deserves to be called 'immensely rich' because it doesn't just require the reference between individual concepts and their referents to fit, but it needs to place these concepts into a set of relations to other concepts. Such a set of concepts that are all innately placed in relations to (even just a few) other concepts would be a very large cognitive load. I take it that Fodor, or any defender of AC, would have to give either an explanation of how this innate knowledge is plausible, or give an argument that this actually isn't a necessary consequence of an innate language of thought. Unless Fodor offers an answer to this, an instance of the first problematic point made above applies. Thus, the Holist challenge arises:

Holist Explain how the richness of relations between concepts can be innate.

Note that the Holist challenge is not in itself a challenge for the innateness claim, but for the strong version of Nativism we see in Fodor's earlier work (Fodor, 1975, 1981). Sellars doesn't oppose the idea of a (restricted) class of innate concepts (see e.g. Sellars, 1980, §106), but does not accept Fodor's account of how such innate concepts can become known to a concept user. Specifically, he opposes the proposal that they can become fully operational simply through the first event of contact with – or a simple triggering by – the kind of object that the concept refers to, or is about. This latter idea, however, is at the heart of AC. We will discuss Sellars's alternative to the idea of triggering a concept in more detail below.

5.3.3 The Relational challenge

A final Sellarsian challenge for AC lies in the theory of meaning that traditionally came with it: it is a problem for relational accounts of meaning in general. Consider an example such as "'Rot' means 'red'", a classical case of giving the meaning of a term in a foreign language. For a relational account of meaning, this statement works as a meaning-giving statement because it makes the meaning-relation clear: the relation that holds between 'red' and red things is the same relation as the one that holds between 'rot' and red things: Both words *mean* the same things, so they have the

[5] For example, one might say that having the concept green is sufficiently well supported iff one can infer from 'this is green' 'this is not red', 'this is not blue', and 'this is not yellow', without being able to judge 'this is not magenta'. Even for an inferentialist, the inferences have to stop somewhere. This certainly raises the problem of the analytic-synthetic distinction (Is the inference to 'this is not turquoise' essential for mastery of the concept GREEN?), which would need to be addressed in a defence of a Sellarsian theory of concepts.

same meaning. At first sight, this works nicely for a wide range of concepts (typically concepts of observable objects and events). In Sellars (1956, §31), he argues that this model runs into difficulties when trying to explain sentences such as "'und' means 'and'", since there is no thing in the world that both words refer to, and it sounds rather opaque to say that both 'und' and 'and' refer to CONJUNCTION (as deVries and Triplett, 2000, p.63, point out) – unless the meaning relation is not between words and things in the world, but between words and Platonic universals. The same goes for abstract concepts like JUSTICE. There seems to be nothing in the world that relates to our concept of justice: an action might be called just, but it cannot be more than an exemplar of what might be called the universal 'Justice'. Let's call this the 'relational challenge' to AC:

Relational Explain how meaning can be a relation when lots of concepts are not in a relation to an observable object or event.

The debate between Platonists and nominalists shall not become the topic of this chapter, so suffice it to say for now that Sellars qualifies as a nominalist, and treats universals with strong suspicion if they are not explained through a linguistic framework, thus eliminating the need for Platonic universals (see Sellars, 1963, 1968, among others). Sellars's own alternative to the relational theory of meaning is the theory of meaning as functional classification (as in Sellars, 1974, p. 421ff.), which capitalises on the relation between the words, and the functions that they play in any given (compared) speech communities, as observable from language use and as can be brought into rule-like form through 'meaning statements'.[6]

If all of these points go through, and Fodor cannot answer these challenges without giving up important parts of his theory, then his position becomes untenable. Without a distinction between 'awareness of' and 'awareness as', Fodor's theory is imprecise and allows too easily for concept possession. Without an answer to the challenge of the holistic nature of systems of concepts, his position is at risk of implausibility because of the conceptual system's alleged innate richness. Finally, without an account of intentionality that can cope with the 'non-relational' nature of abstract and logical concepts, Fodor's philosophical program itself would be incomplete. On the basis of this analysis, we can question the strength of premise 3 of Fodor's Challenge. Until Fodor has given an alternative argument for his conclusion, or until he has sufficiently answered the above challenges, we can regard his position as weakened.

With this picture of Sellars's critique of Fodor's position established, I will proceed by analysing Sellars's own theory of concept learning. I will begin by sketching the prerequisites for his theory, which take the observations made in the critique of AC as the foundation for an alternative framework for concept learning. I will continue with an exposition of the theory of learning in a social frame which Sellars most extensively

[6] I will return to this aspect of Sellars's theory later, in order to now focus on his criticism.

discusses in Sellars (1969, 1974). Following up on this, I will investigate the relation between concepts and words in Sellars's theory, and will offer two possible interpretations of his stance.

5.4 Sellarsian concept learning

Although Sellars makes his original points about language learning, our discussion can be extended to concept learning. Language learning and concept learning are closely related in a Sellarsian theory of concepts: having a concept is understood as the ability to make discriminative judgements about membership in a given category. Learning a word, on the other hand, concerns learning what things or actions should be described, or referred to, by using that word, which also has a discriminative aspect. Therefore, learning a word and learning a concept have a large common basis. I will take this as the starting point for developing Sellars's ideas on concept learning. In this section, I will first present Sellars's alternative to Fodor's premise (3), which makes concept possession conditional on the social framework of a language community and on the community's readiness to ascribe a concept to a given 'new' member. Some themes from the previous section's critique will resurface here as well. Having given the precondition for concept possession, I will present the learning mechanism that Sellars describes for such socially scaffolded learning episodes that lead to fulfilling the possession condition.

5.4.1 Logical spaces

Wilfrid Sellars's philosophy of mind is based, among other things, on the idea that the use of concepts is only possible within the frame of a 'logical space' that was created by a community of language users. It replaces Fodor's possession condition, and promises a non-Augustinian picture of concept learning, as we shall see in what follows. To get a grip on Sellars's alternative to AC, let's begin by looking at the notion of 'logical space', as it is used in Sellars (1956). For this, I will focus on §30 of *Empiricism and the Philosophy of Mind* (EPM), going through the relevant passages step-by-step, adding interpretation and reference points.

> There is a source of the Myth of the Given to which even philosophers who are suspicious of the whole idea of inner episodes can fall prey. This is the fact that when we picture a child – or a carrier of slabs – learning his first language, we, of course, locate the language learner in a structured logical space in which we are at home. (Sellars, 1956, §30)

Sellars thus thinks it's a fallacy to locate the language learner in said space. He isn't yet at home there, that means that he doesn't conceive of the structures we conceive of. He doesn't yet share the conventions we have agreed to in order to talk about the world and the objects in it. Sellars then goes on to say more about just what it is the language learner doesn't yet share with us.

> Thus, we conceive of him as a person (or, at least, a potential person) in a world of physical objects, colored, producing sounds, existing in Space and Time. (Sellars, 1956, §30)

I take this to mean that it is uncontroversial for us that the language learner is part of said world. Yet, the language learner might not be aware of this in the way we are – he will neither have the categories and terms we use to talk about the world nor can we expect him to use classifications that are at least similar to ours. Sellars isn't saying that the person isn't in the same world as we are, but that it has a different status for him.

> But though it is we who are familiar with this logical space, we run the danger, if we are not careful, of picturing the language learner as having ab initio some degree of awareness – "pre-analytic," limited and fragmentary though it may be – of this same logical space. We picture his state as though it were rather like our own when placed in a strange forest on a dark night. (Sellars, 1956, §30)

For Sellars, there is no degree of awareness of the categories and logical relations between them. The 'awareness' he speaks of can roughly be likened to the 'understanding' Augustine attributes to himself as a child, and which Wittgenstein denies plays a part in the process of learning a language. The kinship between Sellars's and Wittgenstein's critique becomes quite visible when comparing §30 of EPM with §32 of the *Philosophical Investigations*.[7]

> In other words, unless we are careful, we can easily take for granted that the process of teaching a child to use a language is that of teaching it to discriminate elements within a logical space of particulars, universals, facts, etc., of which it is already undiscriminatingly aware, and to associate these discriminated elements with verbal symbols. And this mistake is in principle the same whether the logical space of which the child is supposed to have this undiscriminating awareness is conceived by us to be that of physical objects or of private sense contents. (Sellars, 1956, §30)

It would be a mistake to think that the properties that adult language users – with a conceptual system that is possibly attuned to very fine discriminations between properties and even individuals – can conceive of and link to their occurrences in their perceptual experience are as clear or even conceivable for the concept learner. The example of young infants' difficulty to tease apart colour and shape of an object, as researched by Linda Smith and colleagues, comes to mind to strengthen Sellars's observation. Smith and Kemler (1978) observe that

[7] "Someone coming into a strange country will sometimes learn the language of the inhabitants from ostensive definitions that they give him; and he will often have to *guess* the meaning of these definitions; and will guess sometimes right, sometimes wrong. And now, I think, we can say: Augustine described the learning of human language as if the child came into a strange country and did not understand the language of the country; that is, as if it already had a language, only not this one. Or again: as if the child could already *think*, only not yet speak. And 'think' would here mean something like 'talk to itself'." (Wittgenstein, 1953, p. 13e f.)

> (...) young children perceive as integral many dimensional combinations that are separable for adults. (Smith and Kemler, 1978, p. 503)

The 5-year-old children in this study show a burgeoning ability to separate the size and the colour of printed squares after several trials. Younger children, however, perceive the stimuli as holistic and cannot, for instance, sort by size but independently of colour.

One might take the dichotomy of 'awareness of physical objects' versus 'awareness of private sense contents' to designate the divide between Empiricist theories of one extreme, which posit a form of direct perception, and of traditional Empiricism, which takes certain concepts of the sensory world as innate. The former view is usually associated with direct, or naïve, realism of the kind endorsed by Russell (1959) or Putnam (1999) (cf. deVries and Triplett, 2000, p.9f.). Interestingly, the latter position is shared by those traditional Empiricists and by Nativists of some persuasions alike, as we have seen in the discussion of Fodor (1981) in Chapter 2.

In §36, Sellars most famously uses the expression 'logical space', and we should also take it into account here:

> The essential point is that in characterizing an episode or a state as that of knowing, we are not giving an empirical description of that episode or state; we are placing it in the logical space of reasons, of justifying and being able to justify what one says. (Sellars, 1956, §36)

First, note that, for present purposes, I'm placing the emphasis on '*logical space* of reasons' rather than on 'logical *space of reasons*'. When thinking about concept acquisition, and even when thinking about early language acquisition, 'giving and asking for reasons' isn't the primary activity one can use as evidence for mastery of a given word or concept. Second, the quote points us towards the discursive nature of the logical space – it can only properly be called so if those within it are able to talk about it and share their judgments regarding its events, properties, or objects. The emphasis is on *talking* because this is the most versatile form of sharing views and exchanging arguments. It's hard to think of many other ways in which we can give reasons for something we do, or want to defend as a (non-linguistic) statement of a fact (whatever that would amount to) – one might present physical evidence for it (showing the actual fish one has caught, and not just show its size with one's hands).

Sellars thus thinks it's a fallacy to locate the language learner in the logical space. She isn't yet at home there, that means that she doesn't conceive of the structures we conceive of. The point is not that prelingual thinking should not be called 'thinking', since it may very well be directed at the same phenomena, or objects in the world, and may employ similar inferential mechanisms; rather, we shouldn't expect the infant to use the adult's ontology, or adult categories, in thinking about this very same world. She doesn't yet share the conventions we have agreed to in order to talk about the world and the objects in it.

Issuing from this analysis of Sellars's use, the following two descriptions have the purpose of (a) defining what a logical space is and (b) how the idea of a logical space frames the concept-possession condition that Sellars offers as the alternative to Fodor's condition in his premise 3.

Logical Space A Logical Space is a normatively structured system of categories which guides (adult) human conceptual activity.

Logical Space Requirement (LSR) Possessing a concept means having learned the concept's place in the language community's Logical Space.

In the following section of this chapter, I will flesh out in more detail how the LSR works as the culmination point of concept learning. Note for now that there can be many different instances of logical spaces, and there needn't necessarily be one explanatorily or ontologically primary such space. Another issue worth noting is that the name 'logical space' isn't meant to refer to the human capacities of logical reasoning. The possibility that these capacities develop independently of the acquisition of concepts in the logical space shall not be a topic of the present investigation, despite being a highly relevant to developmental psychology and the cognitive sciences.[8] One important consequence of this view is that an essential part of the mental life of human beings is social in its origin – the decisive building blocks of (at least) propositional thought, concepts, are social, cultural accomplishments.

5.4.2 Learning without 'knowing the rules'

Sellars advances a rudimental theory of word learning that is built on his 'verbal behaviourist' theory of meaning and of thought generally. Its main component is the idea that the learning of a word is based on pattern-governed behaviour that is not yet directly guided by the rules of the language. The notions 'pattern' and 'rule' deserve special attention in this context. I will begin by discussing rules. Rules are "general statements concerning what ought or ought not to be done or to be the case, or to be permissible or not permissible" (Sellars, 1969, p. 507), and in the learning episodes that Sellars discusses, they are the means by which a teacher instructs a learner.

Now, a defender of Fodor's view might become quite suspicious of the notion of learning through rules. It might remind her of Fodor's argument in support of AC, as rehearsed above. As a matter of fact, Sellars is aware of this danger, and thus makes a decisive distinction between two kinds of rules. Here's the first kind of thought one might have regarding rules: in order to follow a rule, one must be able to recognise situations in which the rules apply - one needs to have the concepts pertaining to a situation, and one needs to be able to infer the correct response to be taken in that situation (cf. Sellars, 1969, p. 507f.). Sellars calls rules of this kind 'rules of action', or

[8] In his reply to my paper at the AGPC 2011, Michiel van Lambalgen showed me the importance of stressing this point.

‘ought-to-do’s. They are, however, not at play in the *learning* of new concepts, since that would lead to the circular position that, in order to learn concept C, you have to be able to recognise a situation as pertaining to the use of C/being appropriate for the application of C – you have to have C. That means that Fodor and Wittgenstein are conceiving of the rules for learning a predicate as rules of action, which explains why Fodor gives this argument against Wittgenstein, and why he would favour AC.

So, since the route of using rules of action for language learning is not open for Sellars, a second kind of rule is needed to explain these learning episodes: ‘ought-to-be’s, or ‘rules of criticism’. Their name indicates that they don't have the function of prescribing an action: instead, they prescribe a goal state. This second kind of rule can be obeyed even without having the concepts; one can even formulate rules which don't need to be obeyed by their subject, such as “Clock chimes ought to strike on the quarter hour.” (Sellars, 1969, p. 508)
The utterance of a word in reaction to seeing an object is taught along rules like

> *(Ceteris paribus)* one ought to respond to red objects in sunlight by uttering or being disposed to utter ‘this is red’ (Sellars, 1969, p. 511)

The learner is not the one faced with such rules of criticism, but she is subjected to them through the teacher, who follows the rule. So, the learner doesn't need to have any concepts at all when starting to learn the concepts of her community's language. She is guided by her trainer's judgement of her performance in responding verbally to stimuli in her surroundings.

The second main notion in Sellars's theory of concept learning is “pattern-governed behaviour”, which is closely related to the account of rules that I have just introduced.

> Roughly it is the concept of behavior which exhibits a pattern, not because it is brought about by the intention that it exhibit this pattern, but because the propensity to emit behavior of the pattern has been selectively reinforced, and the propensity to emit behavior which does not conform to this pattern selectively extinguished. (Sellars, 1974, p. 423)

Successful teaching by rules of criticism results in the learner showing behaviour patterns upon encountering a given stimulus, as in the case of learning the concept RED: the learner either utters or is disposed to utter ‘This is red’ when presented with a red object. In one sense, this is a pattern because it's more than just the (in this instance) adjective describing the object, but a whole sentence. So, learning about perceptible features brings with it that one also learns basic grammatical structures. A word is not learned on its own, but always in a proper syntactic embedding. In another sense it is a pattern because the teacher sets up a correlation between a given feature of the surroundings, a response or disposition to respond, and an appraisal by the trainer (if the trainer is around). This feature of the pattern-governed behaviour account is important insofar as it sustains Sellars's broader philosophical aim to argue that some linguistic behaviours are not actions in the conduct sense (they are not a piece of behaviour one decides to show), but rather they are automated, and thus there is no need to see thought as something distinct from linguistic behaviour.[9]

The idea of intra-linguistic behaviour patterns is clearly conducive to the type of conceptual holism that Sellars's theory of concepts is built upon, and to the inferentialist position that the 'game of giving and asking for reasons' requires. When we use one word instead of another one to describe a given object, or situation, we commit ourselves to potentially saying a lot of other things about it – when we call something an animal, we can be expected to also endorse that this thing is a living organism (or was so until it died), has come into this world by a certain biological process, and many other things that might specifically apply to the animal species in question. Thus, when using the word 'animal', we demonstrate the ability to judge a thing to fall under a concept, and to link it with a set of other concepts. To be credited with correctly applying the concept, we need to potentially be able to apply several other concepts that are related to the first one. This kind of conceptual holism is inherent in the Sellarsian theory of concepts, and brings its own set of problems into the discussion. I will bracket these problems for now.

Another consequence of this story of concept learning, combined with Sellars's concept-possession criterion, the Logical Space Requirement, is that the process of learning a concept is only finished when one is in a position of being a trainer oneself – having a concept means being part of the logical space of reasons, and this means knowing the rules of action that underlie the rules of criticism one was the subject of. So, in a sense, concept learning in the Sellarsian way is not a piecemeal activity of learning one concept at a time; rather, the learning process for one concept is only finished once a potentially large group of concepts is learnt. Also, Sellarsian concept learning is not based on learning about one type of concepts (e.g. colour concepts) first, and only later learning more elaborate concepts (e.g. linguistic vocabulary, theory-of-mind vocabulary):

> The language learner gropes in all these dimensions simultaneously. And each level of achievement is more accurately pictured as a falling of things belonging to different dimensions into place, rather than an addition of a new story to a building. (Sellars, 1974, p. 425)

5.4.3 The advantages and the problem of Sellars-type learning

To summarise what we have covered so far, I want to demonstrate that Sellars can accommodate all of the challenges that hindered Fodor's position. Following up on

[9] This is a very short, and perhaps puzzling presentation of Sellars's views, and it is not at the centre of the present investigation, so suffice it to give a very short explanation of how this fits into his theory: Sellars (1968, 1969) argues that dualists of a Cartesian tradition hold the view that some thoughts are actions, and some are merely actualities – things that happen without being chosen to happen by the thinker. But all speech acts are actions. So, we cannot base our understanding of thought on our understanding of verbal behaviour. Now, with the idea of pattern-governed behaviour, Sellars has given a reason to see some speech acts as actions and some as actualities, and thus has opened the terrain for exactly this explanation that the Dualist thought ruled out.

this, I will raise one problem for Sellars, which will concern us for the rest of this chapter.

First, Sellars deals with the Holist challenge by conceiving of concept use as placing oneself in a position of making inferences guided by norms of use. This requires the acknowledgment that every concept stands in an inferential relation with a potentially large class of concepts. The exact nature of the holism Sellars would accept doesn't need to be the strong position that every concept stands in some relation to every other concept. A certain form of 'molecularism' can potentially do the job as well – instead of 'every concept stands in a direct relation to every other concept', the claim might be 'every concept stands in a relation to every other concept via a proxy of n other concepts'. The concepts GREEN and TRIANGULAR could be related because they both apply to objects, thus having the proxy concept OBJECT linking them in a conceptual system.

Second, the Awareness challenge is met by Sellars as well, since he tells a learning story that places the learner in a position of being aware of any number of objects or things around her, depending on objectively measurable conditions, without assuming that she has any awareness of those objects as whatever they are conceptualised as by her peers.

Third, the Relational challenge is met because Sellars gives an alternative theory of meaning – meaning as functional classification (cf. Sellars, 1968, 1974). Expressing the meaning of a given word or expression isn't seen as putting the word in a relation with an object in the world or with an abstract entity, but as characterising the word as one instance of a linguistic kind, written as a dot-quoted expression "the · x ·".[10] One can roughly characterise Sellars's theory of meaning as functional classification as an early instance of functional-role semantics. It treats a particular word token such as 'sparrow' as an instance of the linguistic kind, or functional role, ·sparrow·, which covers tokens of any languages that have a word for the expression. One has grasped the meaning of ·sparrow· if one uses the word 'sparrow' correctly in one's reactions to perceptual inputs, in one's intra-linguistic inferences, and one's behavioural reactions to tokens of the word.

The Sellarsian theory of concept learning can be seen to do significantly better in cases like learning AND or JUSTICE than AC with its relational theory of meaning. Occurrences of two things as the objects of a learning episode could for instance be framed by a rule of criticism of the form "(ceteris paribus) one ought to bring it about that one utters 'there is a x and a y' when in the presence of a conjunction of an x and a y". Occurrences of only a y would either be chances to use a new rule of criticism ("...utter 'this is a y...'") or to correct the learner if she uttered 'there is a x and a

[10] The device of 'dot-quoting' an expression is important for Sellars's philosophy of language, but perhaps too technical to properly introduce for the purposes of this chapter. Very shortly, a dot-quoted expression is the 'meaning' of a given word in any natural language that has a term that roughly denotes the same thing. A ·this is red· is the meta-linguistic functional classification for any kind of language that can express *that this is red*, e.g. 'das ist rot', 'c'est rouge', 'det är röd', etc.

y'. Learning episodes for abstract concepts like JUSTICE might involve somewhat more complex rules of criticism, but are ultimately of a similar form.

With this short characterisation of the main features of Sellarsian concept learning, I want to turn to a major problem for this kind of learning story. The Logical Space Requirement upon which his theory is based rules out prelinguistic concepts – as it stands, Sellars seems to bar pre-linguistic infants from conceptual knowledge.[11] They don't have the means to engage in discourse about reasons for taking things to be this way or another. How do they learn to do this? As we have seen, Sellars sketches a theory according to which children are trained to conform their linguistic practice to rules until they finally master those rules and become able to think about their own thoughts in the well-ordered way postulated above. But the question still stands: how do prelinguistic children even get to the point of being able to follow such rules? What kind of cognitive apparatus would be required for that? I will take this as the main problem for Sellars's views on concept learning, and will not discuss Fodor's possible rejoinders for now. While Fodor has reacted to 'holist' or 'pragmatist' types of theories of concepts such as Sellars's (Fodor and Lepore, 1992; Fodor, 2008), I will proceed by looking at Christopher Gauker's post-Sellarsian theory of concepts, since Gauker puts forward a bold hypothesis: concepts are words, and any non-linguistic cognitive activity is non-conceptual. I will end this chapter by arguing against Gauker's construal of what concepts are, and will use a strand of developmental research to support an alternative route that post-Sellarsian theorists might take.

5.5 Gauker's view: imagistic cognition before conceptual thought

Philosophers in a broadly Sellarsian tradition have developed his view to mean several things. So, for instance Christopher Gauker's (Gauker, 2005, 2011) thoughts on concepts are inspired by Sellars's work, although he rejects important parts of the 'verbal behaviourist' theory, such as the functionalist theory of meaning (cf. Gauker, 2011, ch. 4). He however retains a version of the thesis that there is an important role for normativity in explaining the meaning of words.[12]

Gauker's view is that having a concept can only mean knowing a word, and prelinguistic evidence for concept use is only evidence for the infants' ability to make similarity judgments across sample sets of category members vs. non-members – as researched in studies like that of Quinn et al. (1993). Gauker would hold that language is not necessary in order to be able to make similarity judgments since those latter judgments are not a form of conceptual activity. In this section, I take a closer look at this position

[11] This might only be an impression, however: there are various places in Sellars's work in which he affirms the possibility of pre-linguistic representational capacities, especially in his later work (Sellars, 1980, 1981b). In a positive account and extension of the Sellarsian picture, an appraisal of his views on representational systems would support the points I am about to make below.

[12] Gauker calls it a 'third conception' in recent talks and in a forthcoming paper, and opposes Brandom as well as Sellars in important respects.

and I raise some objections to it.

5.5.1 Categorisation tasks in early infancy

As already discussed in Chapter 3, the most successful research paradigm in infant cognition is arguably the dishabituation, or familiarisation, technique (see e.g. Murphy, 2002, p. 272f.). Due to the behavioural limitations of young infants, the most telling piece of evidence one can gather when having them participate in experiments is what they look at, and how long they look at it. The principal idea is that longer looking times can be translated into greater interest, or bigger surprise, depending on the situation. Seeing an object, or a type of object, over and over makes it less and less interesting, and infants will look less and less at it, until a new or unexpected object 'catches their eye' again and they look at it much longer. Paul Quinn has used this study design in his research on young infants' categorisation behaviours, and has been criticised by Gauker (2011) for his conclusions from doing so. As a reminder, let me take a look at the general idea of the study and at Gauker's criticism.

Quinn et al. (1993) have researched the reactions of 3-4 month old infants upon being shown images of cats, dogs and birds. After having seen a number of pairs of cat pictures, the infants dishabituated when being presented with a pair of a cat picture and a bird picture. The researchers have taken this as evidence that the infants have formed a category that includes only cats, and their surprise at seeing a pair of a cat and a bird stems from their categorial differences. Gauker objects to this explanation, and offers an alternative explanation which doesn't rely on the formation of categories, but on non-conceptual similarity judgments. A similarity judgment is a thought along the formal lines of a three-place predicate statement:

> 'Object *a* is more similar to object *b* than to object *c*.'

In Gauker's view, this type of judgment is not conceptual because it doesn't need any conceptual framework to function: even though we would describe an infant's thought during the cat-and-human trial phase as "This thing [the bird] is less like this thing [the cat with which it is paired] than any of those other things [the cat it had already been shown] were like it" (Gauker, 2011, p. 172), the thought itself doesn't need to have any referential relation to a category – it is sufficient to assume that the child has some sort of imagistic thought which places the objects in question in a perceptual similarity space in which the objects differ in their positioning along various dimensions. An observer, or a 'cartographer' of this similarity space may well be able to go ahead and group objects therein as belonging to certain categories, but Gauker holds that these categories "may represent nothing psychologically real in the infants at all." (Gauker, 2011, p. 174) Since Quinn et al. cannot use their study to point to even a vague point at which one category in the infant's mind begins and where it ends, Gauker concludes that they should not go ahead and posit conceptual representations. The next step in his argument is that any kind of study in prelinguistic infants' concepts can be explained in a similar way. He holds the view that concepts should be equated with words, and that all non-linguistic cognition should be explained in terms of imagistic thought. To keep the focus of the present investigation clear, I will not go into the 'imagistic cognition' part of his theory, but instead raise a problem for the view that similarity judgments, involving the three-place predicate introduced above, are independent of the use of categories or concepts.

5.5.2 Are similarity judgments conceptual?

In his investigation, Gauker describes that psychological research has usually used the notion of similarity as a two-place predicate of the form 'a is similar to b in respect R' (cf. Gauker, 2011, p. 186). Here, the reference to a specific concept that frames the similarity relation between a and b is obvious. In the case of the 'more similar than' relation, however, Gauker maintains that no reference to a conceptual framework is necessary. He maintains that this kind of similarity "may be *absolute*" (Gauker, 2011, p. 186, my emphasis). However, it is not clear at all why he would want to describe this as an *absolute* relation, especially since he continues by denying that absolute similarity needs to be "similarity *all things considered*" (Gauker, 2011, p. 186). The main example Gauker uses in support of his claim is the comparison of three ambiguously shaped figures a, b and c:

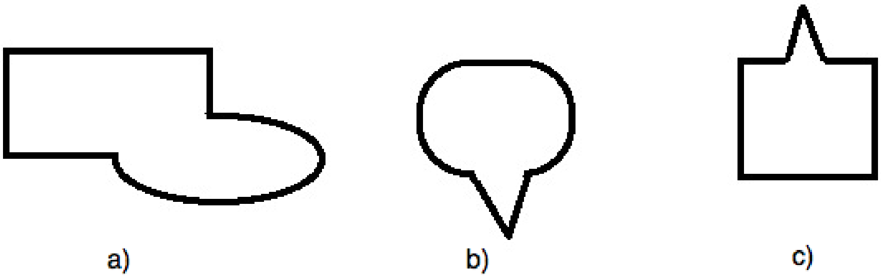

Figure 5.1: Three ambiguous figures a, b, and c.

If we cannot express a clear criterion by which we judge b to be more similar to a than to c, then talk of a conceptual similarity judgment seems far-fetched indeed, or in Gauker's words:

> ... for any general description that readily comes to mind of what a and b have in common that c lacks, that description lets in too much. (Gauker, 2011, p. 190)

To this, I want to reply as follows: first, the force of Gauker's challenge for concept-based three-placed similarity rests on the assumption that concepts indeed are words, which are used to express the exact property that two things share. The thought is that we need to have the words 'conic', 'rectangular', and 'oval' to conceptualise the similarities between a, b and c above. But we only need the words to *describe*, for example, the ways in which *b* is more like *c* than like *a*: both share a conic extension and both are more or less square. But it is certainly conceivable to deny that this is all that there is to forming a concept, and to affirm that one can reliably make similarity judgments without giving descriptions of these judgments. Unless one buys Gauker's conclusion that concepts *are* words, one is free to explain the use of a given concept as the ability to make such (relatively) consistent judgments in similarity tasks. The delineation of the concept as the definite representation of a given category needn't be a part of the concept use. This means that one needn't be able to verbally set strict boundaries for the use of a concept in order to count as having that concept. Just because I will become less sure of my judgments whether a given fantasy animal is more like a fox than a hound dog after I have seen 200 pictures of morphed fox/hound dog pictures, I needn't be worse at the task *because* I don't have the right word to characterise the hybrids. The problem might lie with the vagueness of my conceptual representations of foxes and hound dogs, which makes my judgements less consistent in borderline cases. Nevertheless, drawing sharp lines between different concepts can be a part of the psychologists' tools for explaining a subject's success in a given categorisation task.

Second, his own explanation of non-conceptual similarity judgments, which relies on dimensions of a perceptual similarity space along which the three objects above can be differentiated (cf. Gauker, 2011, p. 191f.), at its most basic, doesn't sound like a non-conceptual explanation. I would take the reliable use of a dimensional differentiation metric to be a tool that supports the formation of thoughts, even without any ability to express such a thought (say, because one lacks the words to *describe* the different kinds of shape at hand). And such tools for thought are exactly what concepts, in the most basic kind of definition, are supposed to be.

All of this is not in principle an argument against Gauker's idea that imagistic cognition in some sense plays an important role in human cognition most generally, since there might be other kinds of cognitive acts that crucially rely on imagistic representations. The point is that Gauker conceives of imagistic cognition as in principle distinct and separate from conceptual thinking, and I want to argue that this latter point is not supported strongly enough by Gauker's ideas on similarity judgments. So, instead of drawing the conclusion, as Gauker does, that we should take Sellars's main insight to be that concepts are like words, I want to present a framework in which this idea is discarded. Instead, I will take the main lesson from Sellars's theory to be that an important part of concept learning and concept-guided activity is inherently social. To support this position, I will proceed by looking at social behaviours that share the social nature of language learning, and crucially depend on the availability of a conceptual apparatus.

5.6 Developmental evidence for prelinguistic concepts

In this section, I want to present some recent results from developmental research that seem like plausible pieces of evidence for prelinguistic concepts. Because of their central role for early cognitive development, I will look at concepts that are needed for understanding social situations, or that might be foundational for social mechanisms of learning, like learning a first language. The focus of this section is on the prelinguistic possession of concepts like PERSON or AGENCY.

5.6.1 Person, agent, gaze-following

To begin with, it has to be noted that the infant concept PERSON may be rather limited (in that it might not fully capture the mental aspect of being a person, like having beliefs and desires) and needn't be identical to that of an adult understanding. For a minimal concept, I would stipulate a definition like

Person An agent that behaves in a goal-directed way and engages with others like her, including the infant.[13]

This is just a rough-and-ready way of defining an essentially non-mentalistic notion of personhood, emphasising that infants perceive other humans as being perceptually like them and that these others are engaged in certain social and communicative practices, which also involve the infant. To support this assumed PERSON concept we are going to discuss, we also need a basic idea of what kind of understanding of agency to attribute to prelinguistic infants. I propose a minimal 'definition-like' description:

Agent An object with the ability for self-started motion towards some goal.[14]

Relating to this, Core Cognition nativists like Carey (2009) argue that human infants show a very early understanding of agency. One marker for this can be gaze-following. Carey cites evidence from children as young as 2.5 months, purporting to show that they are able to follow another person's gaze if certain performance variables are excluded. Then, of course, there's the rich collection of gaze-following data from 9 months of age onwards (for a review, see Brooks and Meltzoff, 2005). On a rich interpretation, the fact that infants follow another person's eyes to see what they see shows that infants understand the concept of goal-directedness as well as the concept of persons as contingently behaving objects with eyes. On a leaner interpretation, gaze-following can be seen as expressing an understanding of behavioural regularities and an understanding of the special role eyes play in behaviour.[15] Without regarding this sort

[13] Note that this definition would not exclude a humanoid robot, or an ape that was brought up by humans. Carey (2009) describes several studies that are linked to the notion of agency in which young infants seem to treat robots in the same way they treat humans. So, clearly the infant concept of PERSON that I posit here is broader than the grown-up one.

[14] Research that supports such a preliminary definition is discussed in Mandler (2004, ch. 5).

[15] See also Farroni et al. (2002) on newborn and 4-month-olds' understanding of the special role of eyes.

of ability as being based in the concepts like PERSON, one might argue, it isn't possible to learn how to engage in more elaborate forms of social interaction with others – one would be barred from fully participating in one's society.

The evidence on the one hand suggests that the infant has understood that the other person is an autonomous agent who initiates her movement herself and often acts goal-directedly in some sense, and that it on the other hand serves as a way of learning from others: connecting gaze-direction, object of attention, and (demonstrative) utterances can be seen as a first step towards learning a first language. This intuition is supported by Brooks and Meltzoff's results (cf. Brooks and Meltzoff, 2005, p. 539f.), and similar claims can be found e.g. in Tomasello and Farrar (1986), among others. Still, one might be sceptical and maintain that all the children do is look somewhere, which doesn't presuppose any concepts on the part of the child. To support the point that infants use concepts in social cognition, I will need to give more evidence, and good reasons for understanding this kind of behaviour as conceptually grounded.

While I am raising concerns about the strength of this evidence, I should also bring Gauker's position with regard to this kind of work to the table. After all, I have been criticising his explanation of prelinguistic cognition, and he might have a strong objection to the line of thought I am following here. But the only thing he has to say about this alternative way of explanation is that other possible paradigms one could explain along that line, like Carey's or Baillargeon's studies in the OBJECT concept (as summarised in Carey, 2009), is that

> that research doesn't even purport to show that children subsume particular objects under various kinds. (Gauker, 2005, p. 291)

However, this raises the fundamental question whether the latter should be the only criterion by which to judge whether something is a concept. My argument against Gauker should deter us from accepting this particular conception of concepts. This is because in the present investigation, using a concept is not just subsuming something under a kind, but displaying certain abilities in acting towards, or on, an object.

To strengthen my point, I have to give more evidence for a development of social concepts of some sort around the age of 9 months. The literature on children's social cognition has greatly increased in the past 30 years, and the vast amount of work on joint attention affords many different strategies.[16] I will discuss the onset of pointing gestures and the phenomenon of social referencing to strengthen my point.

5.6.2 Social referencing

Social referencing is the phenomenon of basing a decision to act in a certain situation on the response of another person. In one study design (Walden and Ogan, 1988), children aged from 6 months to 22 months played with a parent when either one of two toys (a toy Santa or a toy robot) were approaching them from behind a screen.

[16] See, e.g., Tomasello et al. (2005); Eilan et al. (2005)

The interesting question was whether the children would refer to their parents to see how they feel about the new toy, and whether the children would adopt their parent's attitude towards it. The results showed that all age groups referred to their parent, but there were significant differences between the young (6-9 months), middle (10-14 months) and old (14-22 months) groups. The middle group, being most interesting in the present context, reacted on positive and negative parental reactions to the toy alike and either touched the toys extensively (positive) or almost totally refrained from it (negative). The researchers conclude that the evidence suggests that

> the children were using the parental message to construct an interpretation of the event as a whole which was meaningful to them. (Walden and Ogan, 1988, p. 1239)

So, I find it a fair interpretation of the outcomes that the infants, at least in the age range 10-14 months, understood clearly what the different reactions by their parents mean as guidelines for their own behaviour. They have an (at least implicit) understanding of the importance of their parent's acts for themselves, of the shared object of reference – of joining attention – and possibly of the other as 'being like them' in relevant respects. It would be implausible to simply interpret the infants' behaviour as a kind of mimicking of their caregiver's reaction: if that was all that the infants did, one would only observe a mimicking of the adult's facial expression and bodily stance. Yet, the children engage in a deeper way with the adult's feedback, especially in the case of positive feedback. I take it that the inference from 'I see mama is happy' to 'I should go and play with the toy' is a lot more intelligible if we assume some additional beliefs in the child, maybe such as 'Things that mama is happy about are good'. Additionally, the research suggests that the infants make a relation between the caregiver's attitude and the status of the toy – the toy has a causal role in their playing behaviour. This certainly doesn't *prove* that infants employ PERSON in social referencing, but it is suggestive of complex cognitive activity – which is best explained by appeal to conceptual capacities – during such episodes. This all counts in favour of the children applying a concept of personhood, and taking other's behaviour in social settings as a basis for evaluating a situation.

5.6.3 Pointing

As the second main example, consider infant pointing. In a review article, Liszkowski (2008) explains the distinction between referential pointing and symbolic pointing, claiming that the former is a well-observed phenomenon from 12 months of age on, being predated by grasping for objects and by picking up and showing objects to caregivers. Liszkowski claims,

> social-cognitively, communicating by pointing requires an understanding that people attend to things and that one can direct their attention to these. (Liszkowski, 2008, p. 184)

In his own work, he gives evidence for the actual communicative intent behind pointing – infants want the other person to look at what they find interesting; as well as for the referential nature of their pointing – not only directing the other in a certain direction, or to an ambiguous group of things, but to one specific object. I take the evidence to mean that children at the age of 12 months and before the actual onset of language learning use social information and share information, and even points of view, with others. Given the use of the term 'concept' that has guided the investigation so far, this requires conceptual awareness of the other as being able to attend to something and sharing interest in specific situations or objects.

With these two examples from the literature, I have argued that a substantive part of infant social behaviour is best explained with appeal to concepts, against Gauker's view. Just based on this, I however haven't lent the kind of support to Sellars's concept-learning mechanism that it would need to overcome the problem of explaining prelinguistic concept learning. I will give a sketch of a concept-learning mechanism based on social-cognitive abilities at the end of this chapter, as a means of improving on the Sellarsian model introduced earlier in this chapter.

5.6.4 A cut-off point for speaking of concepts?

To take stock of the two post-Sellarsian strategies that we have investigated, it is worth noticing that the main difference between the two – the status of prelinguistic infant behaviour as guided by concepts or not – is an instance of a more general topic in the philosophy of cognitive science: the question whether there is a continuum in representational capacities across species or across major changes in cognitive capacities in one kind of species. This question is relevant to our present investigation of infant concepts because one's answer to it determines one's stance towards the possibility of prelinguistic concepts, and vice versa.

A proponent of a continuum view (such as Barsalou, 2005) could for instance liken the differences in representational capacities and in richness of representations to a mathematical function, like a parabolic function: we start at a very low representational capacity (one might think of Dretske's north-facing bacteria), and from there we see gradual, or even dramatic, rises in it, up to monkeys and great apes, and humans on the continuum.

By comparison, a defender of the discontinuity view would have a bar graph in mind when putting an image to the hypothesis. Many species have a certain level of representational capacity, represented by a smaller or bigger bar. However, no species would have a bigger bar than humans, and there would be a considerable difference when it is compared to the next-biggest bar. While the wider question is of prime philosophical interest, I will focus on the question whether we should think that there is a cut-off point for concept use within the domain of human representational capacities – specifically, whether we should see language as the prime cut-off point in such a division.

Gauker defends the view that language is the cut-off point for concept use, while I have argued that this decision is based on a conception of similarity that doesn't support Gauker's point. While it is reasonable to suppose that language learning transforms human cognition in important ways, it is far from necessary to sever it from all other cognitive activity by reserving the name 'concept' to words. And indeed, one might be tempted to see the issue as a merely terminological question about the use of 'concept'. However, I think that there is more at stake than just terminology, as I see good reasons to include a broader range of higher-cognitive competences in the conceptual part of human cognition. In the spirit of the arguments concerning infant social-cognitive and joint-attentional capacities that I have given, I would like to propose that Gauker's cut-off point isn't useful since it doesn't take into account interesting phenomena of cognitive development, such as the continuities and relations between prelinguistic social cognition and language learning. Based on this, I think that Gauker's proposal runs the danger of not being conducive to – or compatible with – psychological research in the area.

The alternative to Gauker's view that I have proposed in this chapter is to see concepts as those things that are fertile in reasoning and making judgments about given situations – in line with the broad Sellarsian stream that runs through this chapter, one might call this the inferential force of a given mental token. The representational format of this token is not the primary concern, so with this view on concepts, imagistic thought could be considered as conceptual just as well. If we see this as the 'criterion' for being a concept, then it becomes clear that I am more inclined to hold a continuum view of the kind introduced above. Having concepts can come in degrees of 'conceptuality', and some concepts are richer in content or inferential utility than others.

5.7 A second kind of social concept learning mechanism: Tomasello's Social-Pragmatic Theory of Word Learning

After this long investigation into the prospects of post-Sellarsian perspectives on concepts, I want to finish the chapter by connecting my conclusions about Sellars's concept learning model with a contemporary developmental perspective on infant learning – Michael Tomasello's Social-Pragmatic Theory of Word Learning (SPT). My two main conclusions regarding Sellars's model are that it a) cannot explain prelinguistic concept learning without modification and that it b) rightly puts great emphasis on the social aspects of concept learning, which serves as an interesting angle for improving his model. I will first introduce Tomasello's model for word learning, and then connect its premises to the Sellarsian framework of this chapter. I will further show that this model can also serve as a concept-learning mechanism which can overcome Fodor's Challenge by following Sellars's refutation of Fodor's concept-possession condition. Additionally, I will propose that the learning mechanism itself doesn't seem to be based on the HF model, which adds further support to Tomasello's model, which I take to be a good sign for the prospects of concept learning in the face of Fodor's Challenge.

5.7.1 The Social-Pragmatic Theory of Word Learning

It is fair to characterise Tomasello's proposal as a consequence of the successes in joint attention research, in which his work is central. As I will show now, he regards the social-cognitive side of learning as fundamental.

Tomasello bases his theory on two particular kinds of constraints which are thought to guide children's linguistic development:

> (1) the structured social world into which children are born—full of scripts, routines, social games, and other patterned cultural interactions; and (2) children's social-cognitive capacities for tuning into and participating in this structured social world—especially joint attention and intention-reading (with the resulting cultural learning). (Tomasello, 2003, p. 87)

Both of these constraints are integral for the learning process. The first describes the conventional framework within which a large class of concepts is learnt. The second describes the psychological framework within which children are able to process learning in a social context. We have already encountered both of these constraints above, so I shall leave it at this short presentation.

Apart from these constraints, Tomasello endorses a third important claim about the role of communicative intentions in learning words:

> Learning the communicative significance of an individual word consists in the child first discerning the adult's overall communicative intention in making the utterance, and then identifying the specific functional role this word is playing in the communicative intention as a whole. (Tomasello, 2003, p. 89)

Tomasello is an ardent defender of the idea that young infants can understand other people's communicative intentions, and also their beliefs and desires more generally (cf. Tomasello et al., 2005). This is a hotly debated issue in social cognition research, which I won't go into at this point. However, I take it that a weaker version of this claim would also support Tomasello's learning model. What matters is that infants are able to place new words in a framework of words, social cues, and some kind of shared knowledge that helps them in relating the new words to relevant objects or aspects of their environment.

We can extend this model from word learning to concept learning because it is conceivable and plausible to use the same kind of mechanism. As an example, consider learning PLEXIGLASS by playing with a piece of it and exploring its features with a caregiver. Activities could include 'hiding' behind it, touching it, throwing a ball at it, or other playful things. This would involve joining attention to features of the material, interacting by taking turns in handling or experimenting with it, and the like.

5.7.2 Examining the fit of SPT and Sellars's model

Given that Tomasello's work is largely inspired by Wittgensteinian ideas, it seems natural to expect a good connection to Sellars's work as well. I have identified two main elements in Sellars's concept-learning model: the Logical Space Requirement (*LSR*) and the rule-governed process of learning. We encounter the latter again in Tomasello's first constraint, as given above. The similarity to Sellars's views on rules of action and rules of criticism is striking because the rules and conventions Tomasello evokes are prime examples of norms that structure children's lives without them explicitly knowing or following them.

However, we had earlier identified the LSR as the main issue Sellars faces in his explanation of concept learning, since it precludes infant concept possession. It would have to be weakened to become compatible to a developmental model of concept learning. A weaker concept-possession condition would still place an emphasis on the role of the concepts within a society's logical space, but won't demand that a learner have such a firm grasp on the concepts' role in the logical space. In that system, you could learn PLEXIGLASS without neither having the word 'plexiglass' nor knowing what differentiates plexiglass from glass or from obsidian.

So, I conclude that Tomasello's SPT and Sellars's model are in principle compatible, provided we formulate a weaker concept-possession criterion.

5.7.3 Standing up against Fodor's Challenge

Since Tomasello can be interpreted as agreeing with the Sellarsian critique of AC – which he criticises as the 'mapping metaphor' (cf. Tomasello, 2003, ch. 1) – and since his theory is consistent with a weaker version of the *LSR*, I take it that it has the same prospects as Sellars's model by refuting a crucial aspect of Fodor's Challenge.

With regard to the empirical premise of the challenge, and the possibility of redescribing SPT as an instance of the HF model, I want to note the following. SPT cannot be interpreted as an instance of the HF model without significant losses. The first reason for this is that the content of a to-be-learnt concept — say, PLEXIGLASS – isn't fixed by an individual's act of focusing on the piece of plexiglass, but it is fixed by joining attention, and by having one's attention guided by somebody else.

Second, in my understanding of SPT, the learning is not based on a hypothesis about PLEXIGLASS except probably at a later point in the process. For this to come about, the learner first has to form a hypothesis about the relation between their play partner's behaviour and the object they are attending to (the piece of plexiglass). In such a hypothesis, PLEXIGLASS is plausibly neither present nor necessary, as an indexical placeholder like THIS will be much more readily available.

To conclude, Tomasello's Social-Pragmatic Theory of concept learning shows some great promise as a concept-learning mechanism, provided its philosophical foundations in the Wittgensteinian/Sellarsian tradition are able to support it in an extended philosophical investigation.

5.8 Conclusion

We started this chapter by looking at the Augustinian Conception and its place within Fodor's theory of concepts, identifying it as part of the framework for Fodor's Challenge. AC shows clearly in Fodor's possession condition for concepts, and I have argued that Sellars's criticism of AC generally can be linked to the specific point of reflecting this possession condition. In its stead, we have explored Sellars's notion of 'Logical Space' and the theory of learning that accompanies it. Its main features are the holistic nature of conceptual frameworks, the focus on learning patterns of behaviour, the distinction between following rules and being governed by rules, and the social dimension of teaching concept use.

Having identified the main features of his account, we have also found an inherent problem of this account of concept learning: it doesn't have the resources to explain prelinguistic concept learning. One possible response to this, which we have found in Gauker's account of concepts as words, is to deny that there are prelinguistic concepts. I have given two arguments against this view: first, Gauker's alternative explanatory framework, in which he treats similarity judgments as non-conceptual, doesn't actually exclude the necessity of concepts for such judgments. Second, a look at prelinguistic social cognitive behaviours suggests the availability of some concepts before language learning, thus affording an alternative route to tackle Sellars's problem regarding prelinguistic cognition. Instead of capitalising on the word-centred nature of Sellars's proposal, I propose to centre on the social dimension of his view on concept learning. The preliminary conclusion is that Sellars's framework of concept learning is incomplete as it stands, due to the problem and due to the developmentally inspired argument regarding prelinguistic concepts. To amend it, one would first have to weaken the Logical Space Requirement, since it is the source of Sellars's problem.

A positive story about the compatibility of the developmental view of prelinguistic social concepts to Sellars's theory of concept learning as rule-governed activity in the logical space of reasons would have to play on the similarities between the social mechanisms in prelinguistic cognition, possibly including emotional development as well (cf. Hobson, 2002), and the social mechanisms of word learning, and it would have to tell a story about non-social (e.g. physical) concepts in early infancy. In my short discussion of Tomasello's *SPT*, I offer one concept-learning mechanism that can potentially play this role.

With these conclusions drawn, I will turn to another set of criticisms of Fodor (2008) and Fodor's stance on the relation between concept acquisition and concept learning: the arguments in Margolis and Laurence (2011).

Chapter 6

Learning concepts with sustaining mechanisms

6.1 Introduction

Stephen Laurence and Eric Margolis have made a series of important contributions to the literature on concepts in recent years (Laurence and Margolis, 1999, 2012). They have written extensively on Fodor's views on concepts (Margolis, 1998; Laurence and Margolis, 2002; Margolis and Laurence, 2011) and on the issue of concept nativism more generally (Margolis and Laurence, 2013). In setting out their own theory of concept learning, they challenge Fodor's framework for concept acquisition and give reasons to carve the territory of acquisition and learning in an alternative way. I shall first present their recent arguments against Fodor's Challenge, and closely scrutinise the 'syndrome-based sustaining mechanisms' model (henceforth SBSM model) which they propose as the model for learning primitive concepts. This will be embedded in a discussion of the criteria for concept learning, and of the range of possible mechanisms that could fill the gap between Fodor's two extremes.

As discussed in chapter 2, Fodor's argument against concept learning has changed over the course of his writings. Margolis and Laurence have taken this into account and have changed their argument in support of concept learning correspondingly. In Laurence and Margolis (2002), they isolate the structure of concepts as the main topic in relation to Fodor's argument: whether a concept can be learnt depends upon its structure, or its lack thereof – whether it is complex or primitive. Back then, Fodor argued that structured/complex concepts, which are composed from primitives, can be learnt, whereas lexical concepts, which are unstructured, cannot. In their most recent engagement with Fodor, Margolis and Laurence (2011) have followed Fodor's move to avoid questions of conceptual structure and have instead turned the focus on the question of HF model's status as the only possible concept-learning mechanism. It shall be noted, however, that in the long run, the challenge for theories of concept learning remains the same: giving an argument that, or a model how, a concept that wasn't available to a person before a given rationally guided process can be available to them afterwards. The difference lies in the focus; for the earlier Fodorian argument, a model for learning *primitive* concepts was needed to successfully counter Fodor's Challenge, whereas since Fodor (2008), any concept regardless of structure faces his challenge.

I will proceed with the following steps. First, I will rehearse Margolis and Laurence's current summary of Fodor's argument and I will briefly introduce the replies Margolis and Laurence give to Fodor. Their SBSM model successfully defeats Fodor's position, as I will argue. The discussion of SBSMs and their problems will lead to considering another aspect of Fodor's framework for theories of concept learning: the dichotomy 'brute-causal/rational acquisition'. Rejecting Fodor's position on this, I will present two examples in support of the view that the domain of learning is larger than Fodor's argument would have us think; including at least syndrome-based sustaining mechanisms and the Pask device. I will use the criteria for concept learning that Margolis and Laurence (2011) propose, as a set of sufficient conditions for a concept-learning device. These criteria will additionally give us a surprising result: very simple learning systems such as the Pask device fulfil the criteria; it turns out that syndrome-based sustaining mechanisms are not needed to explain concept learning.

6.2 Margolis and Laurence's response to Fodor

6.2.1 The simple Fodor puzzle

As I have laid out in Chapter 2, Jerry Fodor has changed his argument against concept learning with his most recent position in Fodor (2008). Margolis and Laurence (2011) reconstruct the most recent argument in the following way:

> 1. Concepts (whether primitive or complex) cannot be learned via hypothesis testing.
> 2. There is no other way that a concept could be learned.
> 3. Therefore, concepts can't be learned.
>
> (Margolis and Laurence, 2011, p. 509)

This very short version of the argument captures the essence of what they want to take issue with, which amounts to rejecting each of the points (1) and (2).

6.2.2 Arguments for complex concept learning

In their treatment of complex concepts, Laurence and Margolis provide a wide range of replies to Fodor. First, they propose alternative conceptions of hypothesis-testing-and-formation which don't require the availability of the learnt concepts for that process (cf. Margolis and Laurence, 2011, p. 511f.). This amounts to rejecting Fodor's first premise.

Furthermore, they explore the possibility of learning through hypothesis testing without having the concepts to form the hypothesis. It is possible, they claim, to construe the process of hypothesising as one in which a concept is learnt before the concept is used to confirm the hypothesis. The stage of forming the right hypothesis for the first time may for instance be seen as the moment in which we can say 'the concept is learnt': the moment in which the biconditional formulation of the hypothesis is deemed as correct, and the concept thus deemed as worthy of keeping and using to henceforth referring to 'that thing' comes *after* the concept has already been learnt. This is just one of the several connected proposals which Margolis and Laurence discuss. The arguments in this vein are original and thought provoking, but since I want to focus on the possibility of the addition of new primitive concepts to a conceptual system, I will not discuss them here.

In their discussion of a second way out of Fodor's paradox for complex concepts, they start by considering an intuition that is relevant to my own investigation and which I will put to work later in this chapter, namely that

> there is a considerable amount of logical space between brute-causal processes and explicit hypothesis testing. (Margolis and Laurence, 2011, p. 518)

Intending to fill this gap in logical space, Margolis and Laurence offer three mechanisms that don't rely on the HF model, but nonetheless would count as concept learning (thus rejecting Fodor's premise 2): perceptual learning (similar to Goldstone's ideas, yet presented in a simplified version), learning through communication, and associative learning (cf. Margolis and Laurence, 2011, p. 519f.). I will leave learning through communication and associative learning aside for the time being and instead give a brief run-through of perceptual learning, as conceived of by Margolis and Laurence.

Suppose you are on a walk through the park and happen to see a group of swans swimming in the pond, one of which happening to be black. Since you have never seen a black swan before, you can learn the concept BLACK SWAN from this one experience: your perception opens up a new way of combining two concepts that you hadn't thought of combining before, and so this is a legitimate case of learning, far removed from the brute-causal space of non-learning (cf. Margolis and Laurence, 2011, p. 519). Fodor might reject perceptual learning in two different ways, they go on to explain:

> (1) He can accept that these are examples of concept acquisition that don't involve hypothesis testing, yet go on to deny that they count as genuine examples of learning. Alternatively, (2) he can insist that, despite appearances, our examples covertly involve hypothesis testing after all. (Margolis and Laurence, 2011, p. 520)

Margolis and Laurence go on to argue that both of these strategies don't harm their present proposal. They are, however, leaving out a third option that Fodor has available, and which he himself mentions in Fodor (2008): the 'mere' combination of two concepts that haven't formed a conjunction in thought yet doesn't count as learning a *new* concept (both concepts, BLACK and SWAN, are available to form the new complex concept, after all). At best, the new thing happening in your mind when forming BLACK SWAN upon first seeing such a bird would be the formation of a new belief ('There exist such things as swans of black colour', or something of the kind), but calling this kind of (however enriching) new thought a new concept would be putting the horse before the cart. Concepts aren't beliefs, they are constituents of beliefs, and cannot be constituted by beliefs, as Fodor emphasises (cf. Fodor, 2008, p. 139).

As we have already seen in chapter 3, there is more to cognitive psychological accounts of perceptual learning than Margolis and Laurence (2011) discuss here. I will leave the topic at this point and turn to the main focus of the present investigation: Margolis and Laurence's thoughts on primitive concept learning.

6.2.3 The argument for primitive concept learning

Margolis and Laurence have their own idea of how previously unknown primitive concepts can be learnt. Margolis (1998) introduces his ideas with the example of natural-kind concepts, which he assumes to be primitive. As we shall see, the main examples for natural-kind concepts, animal kinds, can be seen as natural kinds even in a generic interpretation of natural kinds as kinds of things that occur naturally in the world, or some other folk-biological interpretation. I will focus the presentation on Margolis and Laurence (2011), will discuss the proposal's merits, and will voice some problems with the syndrome-based sustaining mechanisms model.

Here is a very short exposition of the argument for the assumption of sustaining mechanisms for concept acquisition.

1. Instead of looking at the relations between concepts, we should look at the relations between concepts and their referents for understanding concept learning (cf. Margolis and Laurence, 2011, p. 523).

2. If there are stable connections between referents and concepts, then these are dependent on a psychological mechanism that ensures these connections, and we shall call these 'sustaining mechanisms' (cf. Margolis and Laurence, 2011, p. 523).

3. While many concepts are innate, not all are (cf. Margolis and Laurence, 2011, p. 537f.).

4. For non-innate concepts, we do have the prerequisites to build the sustaining mechanisms that are required for having the concepts (cf. Margolis and Laurence, 2011, p. 524).

5. Constructing new sustaining mechanisms leads to learning new concepts (cf. Margolis and Laurence, 2011, p. 523).

I will go through these steps in what follows, to clarify the proposal and to show where Margolis and Laurence leave the territory of assumptions Fodor can accept. Right at the beginning of the list of these assumptions, there is the commitment to an informational theory of reference. Specifically, Margolis and Laurence (2011) rely on Fodor's asymmetric-dependence theory, which they judge to be the best candidate around (cf. Margolis and Laurence, 2011, p. 522). Asymmetric dependence theory is designed to give a satisfactory solution to the Disjunction Problem for causal/historical theories of reference (cf. Fodor, 1990). Here is a brief statement of the problem. Most causal/historical theories of reference claim that mental representations carry information about the property or object they refer to (cf. Fodor, 1990, p. 57f). Suppose you take a look into your garden at night and see a small animal making its way across the lawn. You cannot see it clearly, but believe that you are seeing the neighbour's cat, and you trigger the concept CAT. However, you were wrong to think that you saw a cat: the animal that crossed your lawn was a skunk. Within a causal/historical framework, this error cannot be accounted for because the reference of a concept is determined by its causal history – by the things that caused it to be tokened. And as Fiona Cowie argues, "if concepts refer to what in fact causes them (or to what they in fact covary with), then, if your concept was caused by (covaries with) a skunk, it must refer to skunks." (Cowie, 1998, p. 236) So, for a causal/historical theory of reference, such a situation necessitates to include skunks in dark nights as causes for triggering the concept CAT, which a) extends the reference of CAT to *cat or skunk-in-the-dark*, and b) makes it impossible to erroneously apply a concept if the error was of a reliable nature (you would routinely make the error in the situation at hand, i.e., darkness in your garden). So, in Margolis and Laurence's words, the problem comes down to this:

> If (...) there is nothing more to content than information, we would not have a case of error here at all, but rather a veridical application of a concept expressing the disjunctive property [*cat or skunk*]. (Laurence and Margolis, 1999, p. 61)

Asymmetric-dependence theory can solve the Disjunction Problem because it stipulates a hierarchy of informational, nomological relations between concepts and their referents. In our example, Fodor would hold that *skunk* only triggers CAT in some situations, such as dark nights, because *cat* is the primary property to trigger CAT (for more detail on this, see Fodor, 1990; Cowie, 1998; Laurence and Margolis, 1999). With this model, triggering CAT upon seeing a skunk is an error because the content of CAT is only determined by its relation to cats. I will leave it at this short introduction of asymmetric-dependence theory because the details of the theory of reference underlying a concept-learning mechanism are not important for an explanation of concept learning. The reason for this is as follows.

Learning the CAT from seeing a skunk wouldn't establish the correct causal chain, and learning CAT-OR-SKUNK from that situation wouldn't do, either. Yet, if one wants to understand how the learning itself comes about, questions of the correctness of what is being learnt (i.e. questions regarding the reference of a thought to its supposed referent) are secondary to questions of what it is that *establishes* the reference: as long as a concept is learnt from encountering a skunk, even if it's the concept CAT, a psychologist could investigate the circumstances of the learning episode and the psychologically relevant processes within the learner, and use this information to hypothesise a learning mechanism, or as evidence for an available model of learning.

The next item on this list is the commitment to an explanation of the connection between concepts and their objects that is built on psychological mechanisms. I will give this step close scrutiny in the next section.

The third step of allowing for a set of innate concepts while stressing that some aren't innate is vital for Margolis and Laurence as it is required to justify their model's focus on innate biases and learning mechanisms. As we will see in the next section, a certain number of innate concepts have to be in place to learn concepts like BLUE JAY, ZEBRA and WATER, let alone more advanced concepts like PROTON or VALIDITY (about which I will not give any details, due to their learning depending on a different kind of mechanism than the SBSM I will discuss below). This point is addressed in the fourth point of the argument given above: the human mind is built to learn certain kinds of concepts, and being built this way, it is stocked with some elements which should be seen as innate concepts and probably also innate theories, or so Margolis and Laurence want to argue.

Finally, the fifth point does significant work for the argument, since it is the claim Margolis and Laurence need to make to support their position that concepts can be learnt: by finding out how a sustaining mechanism is learnt, we should get clear on how a concept is learnt. With this short presentation of the individual steps of the argument, I will go ahead and present the components of the syndrome-based sustaining mechanism model of concept learning.

6.2.4 The syndrome-based sustaining mechanism model

What are syndrome-based sustaining mechanisms? According to Margolis and Laurence (2011), they are the kinds of sustaining mechanisms that are at play in learning natural-kind concepts. As such, they account for a large class of learning events. In Margolis (1998), there is a discussion of two further types of mechanisms: theoretical and deference-based sustaining mechanisms. Since Margolis and Laurence (2011) don't discuss them in their reply to Fodor, I will also omit them for now. Starting with 'syndrome', we get the following information from Margolis:

> What I have in mind is a situation where someone, while ignorant of the nature of a kind, nonetheless knows enough contingent information about the kind to reliably discriminate members from non-members without relying upon anyone else's assistance. (Margolis, 1998, p. 355)

By 'syndrome', in the specific case by 'cat-syndrome', Margolis means observable properties of a given thing that enable one to single out cats in a given group of things. Here is a definition of his use of the term in the cat case:

Syndrome "A collection of salient properties that are readily open to inspection, and are reliable indicators that something is a cat." (Margolis, 1998, p. 355)

Margolis focuses on salient properties here, which I take to mean 'observable properties'. If we take his talk of 'contingent information', as in the quote above, seriously, then a syndrome can nonetheless also be based on strictly speaking unobservable properties of a thing while still enabling the learning of a genuinely new natural-kind concept. Suppose that a person doesn't have the concept FOX, but has concepts of cats and dogs. She might, upon first seeing a fox, learn a new concept that is based on the observable properties (red fur, bushy tail, four legs, ...) but also based on relational properties that are inferred from the observable properties, she might judge the new animal to look more like a dog than like a cat. 'Looking like a cat' and 'looking like a dog' wouldn't count as observational properties on most standard accounts, but as relational properties that characterise a fox in terms of its similarity to other animals. Reference to such unobservable, or only observable-by-proxy, properties can also be a part of a syndrome connected to a natural kind. Other, still more theoretical inferences from observational properties might also be part of a syndrome, depending on the theoretical sophistication of a learner's conceptual system. In the FOX case, the learner can come to draw a variety of inferences that solidify their fox-syndrome: their likeness to dogs might prompt her to infer that foxes are mammals, just like dogs. Recognising a special feature of a fox's snout might lead her to conclude that foxes have a good sense of smell.

Now, while this look at the notion 'syndrome' already gives an indication of how the learning mechanism might work, we still need to look at sustaining mechanisms to clarify how the learning process operates.

Upon a first look, the term 'sustaining mechanism' could mean a lot of different things. It's worth the effort to go through a few possibilities before arriving at the details of what Margolis and Laurence want. Here's a first definition from Laurence and Margolis (2002):

> A sustaining mechanism is a mechanism in virtue of which a concept stands in the mind-world relation that a causal theory of content, like Fodor's, takes to be constitutive of content. (Laurence and Margolis, 2002, p. 37)

To start with a first option which Margolis and Laurence *don't* endorse, a sustaining mechanism could be a metaphysical relation between the mental representation and the represented object. I would take Fodor (1998) to hold a view akin to this when he speaks of 'locking on to a property', where the link between the mental representation/concept BLUE JAY and the observed blue jay lies in 'locking on to bluejayhood'. Keeping up with the vocabulary and spirit of his asymmetric-dependence theory, Fodor could qualify this by saying that bluejayhood is locked onto because of the nomic relation between experiences of blue jays and the tokenings of the concept BLUE JAY. This shifts the question to "What are the laws governing blue jay-experiences?", with the problem – acknowledged by Fodor (1998, p. 122) as a problem for atomist informational semantics – that it doesn't sound parsimonious to postulate an infinite number of laws governing the relations between mental representations and things represented. However, the problem with taking this as the interpretation of what a sustaining mechanism is lies elsewhere: the nomic relation secures the reference of a concept, but not the learning success, so it isn't suitable as an instantiation of a sustaining mechanism. What has to go right for *a* concept to be learnt is not the reference relation, but the learning process. Recall that even if the reference is not correct, a concept can still be learnt – as in the CAT/SKUNK example above. What is retained as a concept could be regarded as close to independent of the correct reference of the concept, except that there must be an experience of a given thing that one can attempt to learn a concept of.

Even by the very general standard of the above quote from Laurence and Margolis (2002), one can guess that the 'mechanism' that Laurence and Margolis look for has to do more than explaining the reference – it has to explain the events that make a concept satisfy a referential relation to its object.

To continue with the options Margolis and Laurence don't endorse, a sustaining mechanism could be a physiological device, or chain of physiological events, that brings about a certain learning effect. One could imagine that there is a particular shade of green, for which the eyes and optic nerves have a dedicated detection system that activates the SPECIAL GREEN concept instantly, even in somebody who has never seen any colours, and never even learnt about colours. An even more outlandish example would be that a certain impulse on your big toe is needed to acquire the concept HALOGEN, because only that impulse on your toe will send the right signal into a part of your brain where HALOGEN can be activated (and incidentally every student of chemistry at some point during their studies experiences the kind of tickle you would need to acquire HALOGEN). This model, however, cannot be what Margolis and Laurence have in mind because it is at base just a variant of the triggering account, which doesn't explain concept acquisition as a rational-causal process. Sustaining mechanisms should do their work because of something more than just physical changes, and this something more should relate the physical learning event to psychological changes in a stronger way than just through random-looking causal chains.

Finally, a sustaining mechanism could be a set of beliefs about a given thing that contains beliefs essential to a concept, but which nonetheless aren't constitutive of the concept: a concept, Margolis (1998, p. 352f.) argues, is like a label on a folder, and the contents of the folder are the beliefs held about the thing the concept refers to (more on this idea below). This picture is the one that Margolis and Laurence advocate. To be able to evaluate their model as a candidate for an explanation of concept learning, we need to look at its components first, and then engage with how a sustaining mechanism makes us learn concepts.

The two main components of the system of beliefs and dispositions that makes up syndrome-based sustaining mechanisms in general are the following:

> (1) the person's knowledge of the syndrome of the kind; (2) the person's belief that membership within the kind is determined by possession of an essential property (or set of properties) and that this property is a reliable cause of the syndrome. (Margolis, 1998, 357f.)

So, here we have the notion of a 'syndrome' as introduced above, and the notion of 'essential properties' of a given object that displays some syndrome. The first part is needed for generally prototypical members of a kind, whereas the latter one is needed for what Margolis and Laurence call 'fakes':

> A fake is a case where a syndrome that is a reliable indicator of a particular natural kind is instantiated in an item that isn't a member of the kind. (Margolis, 1998, p. 357)

Let's have a look at an example. For learning the concept BLUE JAY, the following things have to happen: first, the learner must have an experience of a blue jay, through which she can get a representation of some of its observable properties – say, the blue colour, that it has a beak, the overall shape of it. To form a new concept, the blue jay must be recognised as sufficiently different from the other kinds of birds the learner has a concept of. If that condition is met, then the learner will form a new conceptual representation, which Margolis likes to compare to a label for a folder, which is the new placeholder for the object (the blue jay) in future thoughts about this kind of bird. Through its embedding in a syndrome, the new placeholder concept gets linked to a lot of information, directly or indirectly. The direct links are the ones concerning the bird's look and other perceivable features. The less direct features are inferentially linked: 'being a bird' is part of the syndrome, and this supports the inference to 'lays eggs', to give just one example.

Now, suppose there was a fake blue jay that the learner saw: suppose she sees a bird that has been painted black and at first, she judges it to be a mockingbird. The appearance clearly points towards an actual mockingbird, but if the learner ever found out that the bird actually is a paint-covered blue jay, then she would revise her judgment (even if the appearance of the bird never changed back to the 'blue jay syndrome' type). This is because, apart from relying on syndromes, learners also rely on a so-called 'essentialist bias', as argued for by several prominent developmental psychologists (e.g. Keil, 1989; Medin and Ortony, 1989; Gelman, 2003): what a thing is, i.e., what kind it belongs to, is determined by its innards, or its genes, or in any case by something that is 'essential' and not visible; depending on the sophistication of the learner, the explanations might vary. Susan Gelman characterises the essentialist bias as a psychological mechanism based on

> the idea that certain categories, such as 'lion' or 'female', have an underlying reality that cannot be observed directly. (Gelman, 2004, p. 404)

A notable example comes from Keil (1989). When young children are asked whether a place at the coast – which looks just like an island except that it has a connection to the mainland (and thus is a peninsula) – is an island, they will typically decide this because of the presence of certain characteristic features like a sand beach or palm trees. For them, the peninsula is an island until they come to appreciate the defining features of islands for their judgments about them. At that time, something can perfectly look as if it was an island (with palm trees, beaches, and all the characteristic features), but they won't call it that unless it is unconnected to the mainland. This characteristic-to-defining shift, as Keil (1989) calls it, is consistent with the SBSM model, on the assumption that the model allows for a development of the sustaining mechanisms (and Margolis and Laurence don't give any indication that they would object to this); early sustaining mechanisms might rely more heavily on perceptual syndromes, and will only later implement the beliefs about essences in the learning mechanism.

To sum up, here is a short description of the steps that have to happen in learning a given natural-kind concept – for instance the concept BLUE JAY. Through a perceptual encounter with a blue jay, the learner starts to build a sustaining mechanism for a new concept, BLUE JAY. This sustaining mechanism is based on assembling a representation of the blue jay syndrome – a set of perceivable properties that are combined into a unique set of beliefs about the newly discovered bird. Once the syndrome is constructed, the essentialist bias 'secures' the syndrome's integrity, as it were, by linking the new natural-kind concept to the belief that natural kinds have essences, hence that blue jays have essences. This serves as a reassurance of kind membership in the face of deceiving perceptual circumstances, or of conflicting evidence. All of this information, once assembled, forms a 'mental folder' which is uniquely connected to a label for this folder: the concept BLUE JAY. The label serves as the representation that is used to form further thoughts about blue jays. The learning process happens in constructing the sustaining mechanism, but learning has the concept as its end product. In line with the atomistic commitment and the informational theory of reference, the concept, as a label, itself has no structure or content, but it is linked to the content that it is supposed to have (qua concept *of* blue jays) through the mind-world relation to its referent, which gave rise to the learning process. The concept is also linked to the folder contents – beliefs that, in their combination, describe the concept uniquely (i.e. they couldn't refer to any other concept). With this picture set up, I want to discuss the merits and potential problems the SBSM faces.

6.2.5 Things to like about the sustaining mechanism model

There are four main reasons to subscribe to Margolis and Laurence's model:

Compatibility to Fodor's system Atomism and the causal theory of reference are main factors of the SBSM model.

Abandoning Radical Concept Nativism Giving room for concept learning is the most important advantage of the SBSM model.

Modifying the Building-Blocks model Changing the explanatory metaphor helps opening up the range of learning mechanisms.

Saving Cognitive Psychology SBSMs explain the connection between beliefs and concepts in an intuitively satisfying manner.

Compatibility to Fodor's system

One main advantage of Margolis and Laurence in the debate about concept learning is that they can keep many of Fodor's thoughts on concepts in their picture. This is of interest because of the following considerations: the less Fodor has to reject within their proposal for a concept-learning mechanism, the less room is left for evading the argumentative force of such a proposal. Fodor (2008, p. 141f.) acknowledges that Margolis (and, judging from their recent publications, also Laurence) shares many of his central tenets, such as atomism, the referential nature of conceptual content, and the metaphysical story of locking to properties as the foundation of reference. Setting their own view up in this way reduces the risk of begging the question against Fodor – they assume only what he would concede as well, and then take their argument towards their desired conclusion. One would assume that Fodor would then have to accept their argument in this situation. In comparison to the Sellarsian position that I have discussed in the previous chapter, there is decidedly more overlap between the competing views in the Margolis and Laurence/Fodor case.

Abandoning Radical Concept Nativism

While being quite open to Fodor's ideas, Margolis and Laurence dispose of two major shortcomings (or at least, of shortcomings a defender of concept learning would see in Fodor's theory). First, they do away with Radical Concept Nativism. In proposing a family of learning models, the sustaining mechanisms, Margolis and Laurence offer a way to enrich the conceptual framework through an experience-based, rational process. For the purpose of the present investigation, this is indeed an important aim.

Modifying the Building Blocks model

Margolis and Laurence's insistence on the possibility of newly learnt primitive concepts, or in the atomistic/computational language new primitive mental symbols, rules out a guiding aspect of a second usual feature of Fodor's theory: the "Building Blocks Model of Concept Learning" (Margolis and Laurence, 2011, p. 522). This is the idea that a given number of primitive concepts are the basic and sufficient stock from which all human thoughts are constructed. The main point of this model is

> that the expressive power of the conceptual system is fixed given its principles of combination and its innate primitives. (Margolis and Laurence, 2011, p. 530)

Proponents of the Building Blocks Model thus deny the possibility of learning new primitives, and use a number of metaphors to back up this nativist intuition: human thought can be compared to a typewriter that can only write in one typeset, and in one type of characters (an English typewriter will never produce Cyrillic symbols, no matter how hard you try), and whatever human thought can express, it has to be pieced together from what's available at the beginning (from what the typewriter is built to put to paper).

Margolis and Laurence judge this model to be "misguided" (Margolis and Laurence, 2011, p. 522) and have developed a model that allows them to reject this aspect of the Building Blocks model, while still maintaining the atomist idea that not all concepts need to have an internal structure, and that these atoms can be pieced together much like you could build a tower of building blocks. By not situating the meaning of a concept within the atom, but instead treating it as the label for a set of beliefs about the object that the label refers to, they are free to have unstructured 'building blocks' as constituents of thought. I take it that they would replace the Building Blocks metaphor with a Labels metaphor, like the one introduced at the end of chapter 3. One might conceive of forming a thought involving the concepts X, Y, and Z as a process of copying the labels of a given concept's folder. We can think of the copies as printed stickers inscribed with the names of the concepts. The folder carries the beliefs about the concept, while the copy of the label represents the concept within the concept of a given thought. By putting the copied labels of the three concepts X, Y and Z in a certain order, say YXZ, we arrive at a complete thought. We can think of this as sticking a set of stickers labelled X, Y and Z in the order Y - X - Z. A still more modern way of thinking of the metaphor would be to think of a thought as a string of hyperlinks within a text document – the hyperlink (the label) stands in for what is actually referred to by the labelling word, which in this example would be a set of websites. The metaphor ignores certain complexities of thought, e.g. the intricacies of syntax and pragmatics, but then again, that's why it is just a metaphor rather than a formal description.

Saving Cognitive Psychology

With regard to its role in concept learning, Fodor's own recent theorising has led him to attribute all questions of concept acquisition to neuroscience, leaving cognitive psychology out of the picture (cf. Fodor, 1998, 2008). Cognitive psychology is concerned with, among other things, the role of beliefs and desires in explaining human thinking. If Laurence and Margolis's theory of concept learning works, then there is every reason to keep concept learning within the domain of cognitive psychology. Finding out more about the human proclivities towards psychological essentialism, syndrome-based thinking and the theoretical and inferential structures of our categorising judgments are only a subclass of the cognitive-psychological research on concepts.

6.2.6 Problems for the SBSM

While this all sounds good as an answer to Fodor's Challenge, I still see a few problematic aspects in the syndrome-based sustaining mechanisms model that I want to raise and discuss briefly:

Asymmetric dependence The asymmetric dependence theory of reference faces serious challenges.

The essentialist bias There are reasons to deny an essentialist bias, or to prefer alternative explanations of the evidence supporting it.

Fodor's worry SBSM learning is not sufficient for concept learning.

I will focus on *Fodor's worry* for now, and will just give an impression of the problems behind the other two points. After this, I will turn to a second worry that

Fodor has expressed, and will treat this as the starting point for the second main argument of this chapter.

Asymmetric Dependence

There are doubts in the literature about the success of the asymmetric-dependence theory compared with other informational theories of reference. Adams and Aizawa (1997) give an argument aiming to show that Fodor's theory of reference is not fit to explain reference by appeal to distal stimuli, as he would have to for it to be a successful theory. In another line of criticism, Cowie (1998, p. 238, footnote) voices doubts about how a law can depend on another law in the sense that asymmetric-dependence theory requires, and seems to suspect that Fodor would have to make up the dependencies in an ad hoc manner.

The essentialist bias

While the idea that humans have an innate disposition to look for hidden essences in things seems intuitive, there are alternative conceptions. Strevens (2000) argues that the three common versions of psychological essentialism appeal to the wrong kind of explanations of typical experiments investigating essences and judgements about natural kinds. Instead, he proposes to explain the paradigmatic essentialist examples by appeal to causal-kind laws, which can explain category judgments without any appeal to essences even in cases such as zebra-like painted horses and the like. A second diverging position on the topic comes from Malt and Johnson (1992) and Malt (1994), who argue that the essentialist bias has a weaker effect on categorisation behaviour than supposed by proponents of psychological essentialism. Malt and Johnson's reason to argue this way is that, in their studies, they observe a stronger reliance on physical appearance than on function or essence, which would speak against essences as the main determining factor in categorisation.

Fodor's worry

Put simply, Fodor's worry is that the syndrome-based sustaining mechanisms model is not sufficient for concept learning. As Margolis and Laurence (2011) say, there are two parts to this criticism, both of which focus on the role that sustaining mechanisms – or, as Fodor calls them in his discussion of Margolis, 'theories' (cf. Fodor, 2008, p. 143f.) – play in concept learning. Fodor puts the first part as follows, speaking of a theory (sustaining mechanism) A and a concept B:

> The trouble is that 'You can learn (not just acquire) A' and 'Learning A is sufficient for acquiring B' just doesn't imply 'You can learn B' (Fodor, 2008, p. 144)

Fodor's point here is that Margolis and Laurence cannot claim that the concept that you acquire when you learn a sustaining mechanism is necessarily learnt itself:

it is possible that you learn a new sustaining mechanism for the concept BLUE JAY without thereby learning the concept BLUE JAY itself – your acquiring the concept might be coincidental or might not rationally rely on the sustaining mechanism. After all, the sustaining mechanism might be based on coincidences, irrational inferences, or other defects that would impede *learning* a concept, since learning a concept has to be a rational process. Fodor gives the example of misguided scientific theories which nevertheless gave their proponents the concepts that they used to refer to their objects of study, like in the case of ancient Greek astronomy. According to Fodor, some ancient Greek astronomers assumed that "stars were holes in the fabric of the heavens," (Fodor, 2008, p. 144) yet they could still successfully refer to actual stars or think about the stars despite the scientifically dubious theory that they used when theorising about celestial phenomena.

Margolis and Laurence can however counter this worry: by calling sustaining mechanisms 'theories', Fodor is importing more into the learning mechanism than they would have meant, at least in the case of SBSMs. After all, building a syndrome-based sustaining mechanism is primarily concerned with integrating perceptual information into a coherent set of beliefs about things that look like the object for which a concept is supposed to be learnt. This process is certainly subject to the usual defects of the perceptual and cognitive apparatus, such as hallucination, illusion, false generalisations, and others. As a final touch to stabilise the SBSM, the essentialist bias secures the newly learnt concept's integrity by adding a supposed essence to the perceptually and inferentially structured syndrome. By comparison, building a theory of pretty much anything requires a far greater number of inferences, assumptions and hypotheses besides the use of perceptual experience and such (supposedly) rather basic biases such as the essentialist bias. To form an actual theory, one often has to rely on thinking about causation, one often has to propose imperceptible objects or entities, one has to integrate the hypotheses of the new theory with existing theories that might be connected to the new theory, along with many more highly complex cognitive feats. So, a first point in reply to Fodor's first worry is to point out that sustaining mechanisms *aren't* theories in the generally understood sense – being more charitable, they are something like proto-theories which lack many essential features of both folk and scientific theories.

The second part of Fodor's worry is only raised as a footnote, but is picked up by Margolis and Laurence (2011) nonetheless:

> For that matter, you can acquire a concept by *acquiring* a theory (i.e. by acquiring it but not learning it). I'm dropped on my head and thereby acquire the geocentric theory of planetary motion, and thereby become linked to, say, the property of being a planet. (Fodor, 2008, p. 144, footnote)

With this second worry, Fodor shifts the 'coincidence factor' from coming between learning a sustaining mechanism and learning the concept, to learning the sustaining mechanism: he points out the possibility of acquiring a 'theory' brute-causally, which would in turn lead to acquiring a concept.

Margolis and Laurence give a direct and convincing reply to this second worry: Fodor's point doesn't weaken their position because it doesn't address what they want to achieve. They stress that all that is needed to refute Fodor's original argument, as laid out in the beginning of this chapter, is that there is at least one case where concept learning occurs. By showing that there might be cases which *don't* work in the way Margolis and Laurence envision, Fodor doesn't give a reason to reject their model. Also, his choice of examples shouldn't be likened to the paradigm cases in Margolis and Laurence (2011), namely natural-kind concepts:

> The information recorded is relevant information that is acquired through perceptual and cognitive processes that have the function of recording and of organizing it into new representations of natural kinds. (Margolis and Laurence, 2011, p. 527)

To sum up, while the worry Fodor has about the SBSM model is an understandable one given his position, it doesn't damage Margolis and Laurence's project, because to do this, he would have to give a different argument:

> To address our challenge, Fodor must demonstrate that it is impossible to learn a new concept when a syndrome-based sustaining mechanism is acquired in this sort of paradigm case: Fodor's flukey cases are simply irrelevant given a proper understanding of the dialectic. (Margolis and Laurence, 2011, p. 527)

I am inclined to agree with Margolis and Laurence on this: Fodor's worry about the possibility of merely acquiring a SBSM doesn't damage their account. Having reviewed these problems for the SBSM model, my intermediate conclusion is that the model succeeds against Fodor's Challenge, which is a major achievement given the fate of some of the other contenders that I have discussed in the previous chapters. While I think that the objections stated above are most worthy of being discussed, there is a further point which Margolis and Laurence attribute to Fodor, and which opens the discussion of their own ideas on concept learning up to a more general framework to talk about learning. I have put this third Fodorian worry apart from the rest, and I will spend the remainder of this chapter developing the consequences of Laurence and Margolis' response to it.

6.3 Criteria for concept learning

At a conference treating Fodor's views on concept learning, Fodor raised another objection against the SBSM model, which Margolis and Laurence (2011) describe as follows:[1]

> (...) our model should be rejected barring an alternative *definition* of learning (i.e. a definition that addresses the motivations that originally prompted the hypothesis testing analysis and that can serve as a principled guide for identifying when learning occurs). (Margolis and Laurence, 2011, p. 527)

[1] See Niyogi and Snedeker (2005) for the proceedings of the conference symposium.

The point that Margolis and Laurence want to make is the following: Fodor is convinced that the HF model is the only model for concept learning. Based on this conviction, he constructed his challenge. Now, if anybody wanted to go ahead and give another model for concept learning, they would have to give another kind of definition of what concept learning is, from which a new kind of paradigm for concept learning then could be developed.

Margolis and Laurence regard this as an unfair challenge to their project. First, Fodor himself has argued repeatedly that definitions are usually impossible to achieve (Fodor, 1981, 2008). Following him in this, Margolis and Laurence refuse to define what learning is. Second, it is questionable whether a single definition of what learning is can cover the most interesting aspects of the various kinds of learning, not all of which lead to concept learning. Third, a less rigid way of framing the question "What is learning?" would be to give broad criteria of learning, which is the road Margolis and Laurence proceed to take. Dodging the Fodorian point in this last way is the best response, since it helps to get a clearer view on the subject matter while still not setting the – by Fodor's own account – unrealistic standard that would come up with a general definition of learning. As Margolis and Laurence (2011) emphasise, the good thing about a list of criteria is that there are degrees of fulfilling a given set of demands. Rather than a strictly binary answer to the question "Is this a case of learning?", they can serve as a more fine-grained for describing the success of a range of potential learning cases. In this section, I will discuss the criteria that Laurence and Margolis offer for calling a given process *learning* rather than brute-causal acquisition. I want to investigate whether their criteria are sufficiently unifying for the different kinds of mental processes that have been characterised as learning. These criteria are as follows:

Change "learning generally involves a cognitive change as a response to causal interactions with the environment"

Function "learning often implicates a cognitive system that isn't just altered by the environment but [...] has the function to respond as it does"

Content "learning processes are ones that connect the content of an experience with the content of what is learned" (all quotes from Margolis and Laurence, 2011, p. 529)

I will explain these criteria with the help of a unifying example. Consider a group of amateur ornithologists who initiate participants in their blue jay observation meetings by various means that are supposed to make them acquire the concept BLUE JAY. I will present three cases that don't count as learning by the above criteria. The baseline case, which would fulfil none of them is the following: suppose the ornithologists invited new members to their club room, where they dim the lights and let them sit on a chair for an hour. Should BLUE JAY be acquired at all (let's assume it is, for the sake of the example), then it would be a mysterious, non-causal process.

Regarding (Change), Margolis and Laurence are careful to point out that there can be learning without environmental influence, e.g. of a priori truths and relations between them, and that there can be environmental influences without a learning effect on the cognitive system, e.g. bumps to the head that *don't* give you the concept BLUE JAY. So, (Change) alone is not a sufficient criterion for learning, since it excludes some cases of learning while including some cases of non-learning. However, those instances of learning that we look at as paradigm cases usually fulfil the (Change) criterion. To use our example, imagine a group of ornithologists who found a special way of hitting new inductees on the head, which makes them acquire BLUE JAY. There is a blunt causal interaction with the environment in this process, but the cognitive system doesn't have the function to learn by that hit, and the content of BLUE JAY doesn't have any bearing on the hit – two out of three criteria aren't fulfilled.

In the case of (Function), we have a more specific requirement for learning that rules out brute-causal cases: for example, when learning a foreign language, there are cognitive processes at work that are supposed to, for instance, link a foreign word to a known word. In the case of learning BLUE JAY through repeated hits on the head, the cognitive process that in the end might eventuate in learning the concept most likely doesn't have the function of using such inputs to produce such concepts. Despite this, one can think up cases in which a genuinely brute-causal process, like the BLUE-JAY headbump, gets used as a reliable way of learning the concept – think of our ornithologists hitting new inductees in just the right BLUE-JAY way as part of a standing and thoroughly tested tradition. A possible reply to this scenario is to point towards debates in the philosophy of biology: talk of functions can be ambiguous, as Allen (2009) remarks; one dimension of ambiguity is whether the function is teleomentalistic or teleonaturalistic. The former refers to the aim being dependent on mental states, broadly construed, as in the ornithologists' case. The latter refers to the biological aim of a given trait or property, typically cashed out in terms of evolutionary success dependent on having the trait. I would assume that Margolis and Laurence have a teleonaturalistic notion of function in mind, which would exclude functions of the ornithologist-initiation-rite type. Another perspective on (Function) comes from Ruth Millikan's work (e.g. Millikan, 1989). She holds the view that functions are best seen as dependent on a given object's causal history of having a given function. Evolutionarily, the heart's function is to pump blood, so any newborn animal's heart will have that same function, independently of its dispositions and current workings at the time of birth. Significantly, her position allows for the inclusion of the ornithologists' initiation rite as an event having a function. For Millikan, some object or action A has a proper function F if

> A originated as a "reproduction" (to give one example, as a copy, or a copy of a copy) of some prior item or items that, due in part to possession of the properties reproduced, have actually performed F in the past, and A exists because (causally historically because) of this or these performances.

(Millikan, 1989, p. 288)

To apply this to our example, each new initiation ceremony of hitting a bird-watcher just the right way is a reproduction of the first time a bird-watcher of this circle got the concept BLUE JAY by being hit this way. While the quote above talks about items, Millikan writes that "learned behaviors, reasoned behaviors, customs, language devices such as words and syntactic forms, and artifacts" (Millikan, 1989, p. 289) can also have proper functions.

Even when accepting Millikan's notion of proper functions, there will however still be a barrier for the typical cases of brute-causal acquisition which Fodor discusses – the (Content) condition.[2] We can illustrate this by the BLUE-JAY example: suppose that the ornithologists have found out that, for some evolutionarily important reason, eating a special kind of rotten apple will induce a very specific kind of disgust, which triggers the concept BLUE JAY. Thus, they offer this kind of apple to new bird watchers. Here, we have fulfilled (Change) and (Function), but there is no content relation between the apple, or the disgust, and the content of BLUE JAY.

(Content) is central insofar as it states the demand for a meaning-conferring mind-world relation between the experience and the learning. Again, it is a criterion that excludes the standard brute-causal acquisition examples as non-learning, while including all of the examples for learning that Margolis and Laurence have on offer. Seeing a bird and then thinking of the bird as a consequence of the experience has such a content-conferring relation. Equally, hearing a professor give a definition of a new term can lead to thinking about that term, for instance by memorising the definition. In contrast, the experience of an itch in your big toe which precedes your acquiring the concept HALOGEN (which might, as stipulated above, be seen as an example of triggering), even if causing the acquisition, cannot be explained without extra assumptions (like that the nerve connection activated by the itch is evolutionarily bound to acquiring HALOGEN).

These three criteria are a good basic set of rules for the discussion of concept learning. They give a more detailed description of what concept learning is than in premise (1) of Fodor's Challenge. There, Fodor describes learning mechanisms as rational-causal processes, and this is reflected in the criteria quite clearly: the 'causal' is satisfied by (Change) and (Content), since those two are there to explain the connections between the to-be-conceptualised object and the to-be-learnt concept. The 'rational' aspect is expressed by (Function) and (Content), since they are there to describe how an experience, a cognitive mechanism and an end product of a learning episode hang together rationally. Finally, to call it a mechanism requires at its base to have something like (Change) and (Function) doing the heavy lifting in terms of producing a scientific explanation of a psychological event.

[2] Aside from that, Fodor doesn't consider cases like the ornithologists' rite, since he is intent on stressing the accidental nature of his favourite brute-causal cases, and wouldn't intend to make them sound as if they were actually telling us anything about successful concept learning.

The congenial fit between the criteria and between Fodor's conception of concept learning is why I want to use this set of criteria for concept learning to investigate one particular strand of thought in Margolis and Laurence's paper. This is the idea, already raised in chapter 3, that hypothesis testing (as in the HF model) and brute-causal acquisition might not be the sole possible processes of concept acquisition, and that the more interesting possibilities between the two extremes should be grouped with the learning camp rather than the 'causal-fluke' camp.

6.4 Between brute-causal and rational acquisition

For this chapter, I want to extract the following main lesson from Margolis and Laurence (2011):

The false dichotomy The juxtaposition of brute-causal and rational-causal concept acquisition is a false dichotomy, and looking at the space in between yields acquisition models that are more akin to concept learning than to brute-causal acquisition, while still avoiding Fodor's paradox.

We have started discussion of this point in chapter 3, and will now go into more detail on the proposal, using the analytic framework of Margolis and Laurence (2011).

As we have seen, Fodor talks about a large span of potential events of cognitive change, and groups everything that is not hypothesis-formation-and-test under the label 'acquisition' (cf. Fodor, 2008, p. 135):

> *There are all sorts of mind/world interactions that can alter a conceptual repertoire* (...): sensory experience, motor feedback, diet, instruction, first-language acquisition, being hit on the head by a brick, contracting senile dementia, arriving at puberty, moving to California, learning physics, learning Sanskrit, and so forth indefinitely. (Fodor, 2008, p. 131f.)

In relation to this, he quite rightly acknowledges that the dichotomy innate versus learnt doesn't exhaust the possibilities, and that there could be other forms of acquisition than learning (Fodor, 2008, p. 132, FN 2). The lesson he draws from this is that the only possible kind of concept acquisition has to be non-learning, as I have emphasised before.

The first thing to notice here, however, is that there is a difference between cases like the bump to the head and cases like learning a second language or memorising a complex dance move: while the former is a case of arbitrary relations between the cause for acquisition and the acquired content, the latter has a rational (in a generic sense) link between the two. For instance, you read a sentence in Swedish and then look up the translation of the words you don't know yet, intending to commit them to memory. As we have seen, Margolis and Laurence (2011, p. 529) use this property as the third main criterion – (Change) – for classifying an episode as a learning episode.

So, rather than accepting Fodor's position on classifying cases of concept acquisition as learning or brute-causal attainment (that there is no learning, and that what is acquired is gotten brute-causally), this difference makes another position far more attractive. The HF model is not the only model for concept learning, because several other mechanisms fulfil the constraints for models of concept learning. The class of brute-causal acquisition cases is only the end point on a scale that ranges from very sophisticated learning models, such as the HF model and SBSMs, to simple learning mechanisms, such as the Pask device that I first introduced in Chapter 4, before reaching the non-learning end-point of the brute-causal. In the following, I will lead us through this spectrum and will show that the criteria from Section 6.3 warrant the aforementioned conclusion.

6.4.1 The HF model and SBSM learning

To start at the very top of the line in terms of sophisticated learning models, the HF model satisfies all the criteria for concept learning, under the condition that one bases at least some of the hypotheses one forms on contact with the environment. This condition is needed since it figures in the (Change) criterion. (Function) is also satisfied, since the hypothesis-formation mechanism is clearly 'built' for doing just what the organism needs to do in order to acquire new concepts (if it weren't for Fodor's paradox). Finally, (Content) is (ceteris paribus) fulfilled, since experience gives us the evidence we need to form the correct hypothesis to learn a given concept.

Now, how does the SBSM model fare when confronted with the learning criteria? To start off, (Change) is satisfied by SBSM learning, as demonstrated by the examples Margolis and Laurence (2011) give – learning natural-kind concepts based on syndromes is bound to happen through interaction with the environment, which changes the learner's cognitive system. Second, the mechanisms stipulated for SBSM learning display qualities that satisfy (Function). The essentialist bias and the 'syndrome-network' of beliefs both ostensibly have the function to sustain the learning of new concepts, or to interfere in cases where conflicting evidence might change concepts. Finally, the (Content) requirement is fulfilled through the reliance on evidence when constructing a syndrome for a new concept. For every new concept that is learnt, an individual set of beliefs covers the observed properties and the assumed essential properties that make an object a representative of a new concept. So, if all goes well, the content of an experience maps onto the content of a kind syndrome.

This shows that our first contender already fulfils the criteria sufficiently well to be called a model of concept learning.

6.4.2 The Pask device's learning

Further down in the range of possible concept-learning mechanisms, there is Gordon Pask's electrochemical learning-device. Here's a short description of the Pask device, to remind us of what it does. In the experiment described by Cariani (1993), a set of electrodes is put into an aquatic solution of salts. It is then connected with, for instance, a sound-generating device that can be set to produce a set of variously pitched notes. To our ears, the notes are perfectly distinguishable, but the newly set-up Pask device, being just an electric circuit, doesn't have any sensitivity to changes in pitch. This, however, can be changed through training: after getting different kinds of electrical stimuli associated with different sounds, the metallic components of the aquatic solution start to solidify as bridges between various electrodes. Like this, the researchers claim, it is possible to construct an 'ear-like' sense organ for the Pask device, which reliably links a sound to a certain 'neuronal' reaction (i.e. electric activation) within the system. As a sense-organ-like device, it can theoretically be used as an input device for a third element in a sound-processing chain, like an 'answering' sound-producing device: depending on the note that the Pask device detects as the input from the first sound generator, it could for instance produce a complementary chord, or a following melody (a 440Hz sound might trigger an *A minor* chord played until the next input is processed, while a 330Hz sound might trigger an arpeggiated *E minor* pentatonic scale being played).

The Pask device is of great interest for our understanding of concept learning because we can regard it as analogous to a (primitive) neural system. If such a system were a part of our larger human-cognitive system, then we should consider if it could play a role in the process of concept learning – if its very basic activity alone could already be sufficient for it. In this vein, Landy and Goldstone give the device one cautious and one bold interpretation when discussing it as a perceptual-learning device. They first qualify it as an instance of

> existence proofs for how early perceptual devices can be systematically and physically altered by the environment to change their representational capacities. (Landy and Goldstone, 2005, p. 351)

More boldly, they link the Pask device's development to conceptual development:

> It is a device that, when (literally) immersed in the proper environment, develops its own concept of what is relevant. (Goldstone and Landy, 2010, p. 1368)

But in what sense can this act of constructing electric connections in the device be seen as a concept learning process for the primitive concepts SOUND A and SOUND B? Let's try to apply Margolis and Laurence's criteria for concept learning. (Change) is pretty clearly fulfilled, since there is a complete and straightforward causal story to tell about the physical development of the device, and the corresponding changes in reaction to the relevant class of stimuli. (Function) is less straightforward than that. The Pask device gets altered by environmental stimuli, and the nature of the stimuli determines what the device learns. In what sense can there be a function in how the device's parts (such as the aquatic solution of salts) respond to this? The function that can be observed here has indeed more to do with the laws of physics than with the internal constraints of a cognitive system, such as an infant learning her first words. The latter is undeniably far more complex than even a very large Pask device – just compare the numbers of the electrodes and their connections to the number of neurons and synapses involved in the respective learner's brain. Still, within the constraints that hold for each system, there are regularities in the system's changes that cannot fully be explained by just appealing to the kind of environmental influence, but which need to be explained by looking at the laws governing, e.g., electrical currents in aquatic environments. I take Margolis and Laurence to make a similar point about the cognitive changes in a concept learner with a SBSM working when they remark that, in a learning episode,

> [t]he changes presumably are of the sort that our perceptual systems and related belief-fixation mechanisms are designed to subserve. (Margolis and Laurence, 2011, p. 529)

In analogy, if the concentration of metallic salts and the constellation of electrodes are set up in a certain way, then we can observe the kinds of changes that they are designed to subserve, and can correspondingly develop sets of possible learning outcomes which depend on the Pask device's layout as a function of learning.

Furthermore, using the aforementioned account of what constitutes a function, we can find the characteristics of purposeful goings-on in the Pask device along both Allen's and Millikan's lines. If we use Allen's characterisation of teleonaturalistic accounts, we have to find a convincing explanation of how the "more widely-accepted approach [which] treats functional claims in biology as part of the analysis of the capacities of a complex system into various component capacities" (Allen, 2009, section 'Teleonaturalism') fits our case of the Pask device. I propose that we can justify speaking of a teleonaturalistic function at least in the case of a Pask device that is part of a larger system, thus gaining a derived or conditional function. Were a Pask device to become part of a larger system, such as an electronically controlled synthesiser of the kind proposed above (responding to input sounds with corresponding chords or tonal sequences), it would have a teleonaturalistic function within the system (routing the input signals to the right musical outputs). The learning mechanism that brings the device into the state of fulfilling its function of routing the inputs would by implication also have the function of bringing about the result of the device performing its function. And even if one were happy with only relying on the designer's intentions in building the device, one could explain this in terms of a teleosemantic function: the Pask device has a function because its designer intended it to have one. As I have remarked above, Margolis and Laurence probably wouldn't accept this, however.

If we use Millikan's characterisation of proper functions, we can see that the Pask Device *has* a function. Historically, it was built to develop as a primitive sense organ with simple sensory concepts. Any copy of this original device, which stands in the right causal-historical relation to the original device, has the same function, according to Millikan's definition of 'proper function' (cf. Millikan, 1989, p. 288). That means that any device that has been built with the specifications of the original device in mind can be said to have 'inherited' the proper function of the original device. This account differs from the teleomentalistic account insofar as it doesn't refer to the intentions of the designer of the device, but to the first Pask device that fulfilled the role of a primitive sense organ that evolved to represent two kinds of sounds.

It helps to compare the 'function' we have just laid bare with the non-function of the cognitive system's reaction to the concept-acquisition-engendering blow to the head. The difference between the two lies in the fact that the Pask device's physical changes come about in a reliable way, whereas it's not clear why there are the kinds of changes that we stipulate for the hit on the head (apart from head trauma, that is). I conclude from this that (Function) is fulfilled by the Pask device.

Now, what about (Content)? Does the Pask device "connect the content of the experience with the content of what is learned" (Margolis and Laurence, 2011, p. 529)? The answer to this question depends on whether one is partial to calling the Pask device a representational system. If one is (as I am going to be, following Prinz and Barsalou, 2000), then the story goes like this. The electrochemical changes in the system are the effects of the experiences with the set of learning inputs. After the system has been changed (some electrodes are now connected through metallic filaments), the circuit shows discriminative 'behaviour' with regard to the two sounds – the device represents the sound, and so a learning event has taken place. One could call this a mapping between an environmental input and a concept. If one is not inclined to regard the device as a representational device, it would be more difficult; apart from that, since many theories treat concepts as representations, in various senses, to deny the representational nature of a system is to exclude the possibility of it being a conceptual system in the first place.

To conclude, the Pask device fulfils all three criteria to a degree – enough to class it as a concept learner, and not as a brute-causal acquirer.

6.4.3 Could triggering qualify as learning by this measure?

With the perhaps somewhat surprising result that Pask-type learning can be counted as akin to concept learning, one might raise the question: have we opened the door too widely? Perhaps even Fodor's triggering would qualify under these circumstances. If one doesn't want to just go ahead and call every act of acquisition 'learning', then it is important to defuse this objection. So here are my reasons to reject a rational-causal interpretation of triggering. The main difference lies in the role that experience plays for the Pask device's learning compared to a triggering event. For a given set of electrodes in a solution of salts, the learning event is shaped much more strongly by the kinds of input it receives than in a triggering case. With varying signal strength, varying concentrations of salt in the liquid, with changing external conditions (temperature, pressure, ...), a Pask device will evolve in rather varied ways.

For a trigger-able concept to come online, there only needs to be a signal that can activate that concept, and as the example of duck imprinting shows, there are concepts that can be triggered by a very large class of signals (i.e. anything that moves in front of the duckling). This suggests that, formally, triggering doesn't satisfy the (Content) condition (even though it might grow to fulfil the other two conditions), whereas I have argued that Pask-type learning does.

6.4.4 Lessons from applying the learning criteria

Having applied the three criteria to two cases, I conclude that they apply well, with SBSMs scoring more points, as it were, because of their higher degree of cognitive sophistication and their resulting greater ease at fulfilling the criteria. Syndrome-based sustaining mechanisms can, with good reason, be called 'concept learners', and so can the Pask device. Now, if this holds, we might face the temptation to turn against Laurence and Margolis and use the existence of simpler models that fulfil their criteria to undermine their SBSM model. After all, if we can do without it, why should we keep all aspects of the SBSM model, such as the commitment to psychological essentialism and the reliance on Fodor's framework of atomism and asymmetric-dependence theory?

Margolis and Laurence might have the following reply to this idea. A Pask device cannot be a concept learner because it isn't a cognitive device. Remember that they refer to 'cognitive systems' or 'cognitive change' in their criteria. The question is: what work does the classifier 'cognitive' do there? If it means to say that only cognitive systems can learn, then they have to say something about what a cognitive system is, and why e.g. a Pask device wouldn't qualify. One answer they might have is that the Pask device doesn't put concepts together to form thoughts. Another, related answer, might be that the Pask device is not linking the learnt property to other concepts, for a lack of other concepts. Yet another is that the Pask-device cannot generalise from these learning experience, and that the concepts that it learns do not satisfy Evans's generality constraint (cf. Evans, 1982).

But to this set of sceptical points, there is a straightforward reply: their atomistic theory of concepts doesn't itself necessitate any of these things! So bringing these in would amount to bringing in new criteria for concepthood, and possibly also concept learning. These new criteria, if implemented, seem to be important enough to not be ignored when assessing Laurence and Margolis's own proposal, and Laurence and Margolis would have to say a lot more to show that these criteria fit with their SBSM model.

While the temptation to go simpler is strong, there is no reason to succumb to it now since both learning mechanisms are consistent with each other and the burden of proof is to show that they both can't work in concept learning lies with the radical anti-learning theorist. It would be unfair towards the two contenders to expect them both to be able to fulfil the same role, given that they have different target domains of concepts whose learning they aim to explain. Note also that Margolis and Laurence stipulate a large range of alternatives to Fodor's stance, ranging from Perceptual Learning to amended HF-type learning to the SBSM model for explaining natural-kind concept learning. As I want to argue, a set of diverse concept-learning mechanisms is a useful tool against Fodor's argument, and the diversity along the gradient of 'more/less akin to the HF model's maximum-strength learning' will be important for my argument supporting the elimination of the scientific terms 'concept learning' and 'concept-learning mechanism' in the next chapter. But for now, I think that having a multitude of possible concept-learning mechanisms is a good reason to adapt the stance of what I propose to call *Concept Learning Pluralism (CLP)*. As a rough formulation of this position, I propose the following:

Concept Learning Pluralism There is 1) a range of several possible concept-learning mechanisms, 2) whose differences and variance in learning-success we can quantify with Margolis and Laurence's criteria for concept learning.

In summary, what we get from Margolis and Laurence (2011) are a) the SBSM model as a successful reply to Fodor's Challenge and b) a general set of criteria for concept learning that allows us to evaluate other proposals for concept learning, and to categorise them on a continuum as either 'learning' or 'brute-causal acquisition'. Given the success of the Pask device, we see that the bar for learning events is not as high as Fodor (or indeed Laurence and Margolis) would like it to be. As a third point c), we further get a new option of Concept Learning Pluralism as a meta-theory of concept learning.

6.5 Conclusion: Defending the rationality of concept acquisition

With the points made in the previous section, I have established that the notion of rational acquisition can be sustained against Fodor's criticism. While the rationality of the acquisition process doesn't come about through Fodor's single rational acquisition mechanism – the HF model – it can be defended, because the alternative mechanisms, and even Pask-type learning to a degree, fulfil a sensible set of criteria for concept learning. There is the temptation to use Pask-type learning as an alternative to syndrome-based sustaining mechanisms in the way that Margolis and Laurence conceive of them. The only thing that we need from Margolis and Laurence (2011) in the end are the three criteria, since they help us do away with some of the problematic things we identified above, such as the essentialist bias. While both mechanisms are good counterexamples to Fodor's position, using the Pask device as the paradigm for concept learning has the advantage of relieving us of the burden of advanced mental machinery that is at play in SBSMs. As I will argue in the next chapter, accepting both mechanisms as viable only strengthens the anti-Fodorian position because it undermines a basic point that Fodor's argument relies on: that there can only be *a single* kind of concept learning at play in humans.

Chapter 7

Concept learning eliminated

7.1 The state of play

In the preceding chapters, we have reviewed Fodor's Challenge for concept learning and have found that it proves challenging for a wide variety of philosophical and psychological theories of concept learning. Jean Mandler's proposal – Perceptual Meaning Analysis – has been found wanting and in need of amendment as well as further experimental support. Others, such as Robert Goldstone's Perceptual Learning approach and Wilfrid Sellars's pattern-based learning, have been shown to face significant, if manageable, problems. Another promising proposal is Michael Tomasello's Social-Pragmatic Theory, which fundamentally opposes Fodor's views and has its roots in Wittgensteinian ideas. In the end, the most promising proposed concept-learning mechanisms were Susan Carey's Quinian Bootstrapping and Margolis and Laurence's syndrome-based sustaining mechanisms. These two offer the strongest case against Fodor's Challenge, which leads me to conclude that concept learning is possible after all.

Aside from the learning mechanisms that I have investigated in detail, there are several additional ones that I didn't touch upon in this thesis. Among them are the broad range of machine learning, specifically connectionist models; traditional associationist and abstractionist models; and several other, more specialised, mechanisms (such as non-visual perceptual learning, as in musical education).

In Chapter 1, we started this investigation aiming to provide a discussion of theories that can fulfil the acquisition desideratum – a theory of concepts should be able to explain, ontogenetically and phylogenetically, how concepts are acquired and which concepts are acquired. At this point in the investigation, we have found a variety of mechanisms that can fulfil this desideratum.

7.2 The claim

By modest criteria, such as Margolis and Laurence's criteria from Chapter 6, all of the models mentioned above can be called instances of concept learning. But these models vary quite considerably in their specifics. With so much variation in the attempts to explain concept learning, I see some reason to develop a sceptical attitude with regards to the scientific classification 'Concept Learning' (CL).[1] Apart from the previously discussed problems and challenges in the literature on concept learning, there are themes from the philosophy of science that will be able to help us. My first and principal worry that will motivate what follows is that it seems questionable whether the previously discussed mechanisms of concept learning share many scientifically relevant properties that would justify seeing them as instances of a single natural kind Concept-Learning

[1] I propose the following convention for talking about scientific classifications and natural kinds: I will capitalise names for classifications and natural kinds, as in talking about the natural kind 'Gold'. Furthermore, I will use the shorthands 'CL' and 'CLM' interchangeably with the natural kind terms. Aside from this, I will follow the use/mention conventions and will use single quotes when mentioning a natural kind.

Mechanism (CLM). What would happen if this worry turned out to be well founded? In this chapter, I want to argue that this would necessitate the elimination of 'Concept Learning' and 'Concept-Learning Mechanism' as terms of the cognitive sciences. I will call the position 'Concept Learning Eliminativism' (CLE) and will take the following claim as the focus of this chapter's discussion:

CLE The notion Concept Learning should be eliminated from philosophical and psychological theorising about concepts.

In what follows, I will first lay out the argumentative structure on which CLE will rest, along with a discussion of natural kinds and pragmatics in theory construction. With this in place, I will go on to give an argument that supports CLE, based on the two claims that CL is not a natural kind and that there are pragmatic advantages to eliminating CL. While there are a number of objections to CLE, I will show that none of them defeats the position. I will end this chapter by highlighting the advantages and promises of CLE for future research on how concepts are acquired. Before I start this investigation, I want to lay out several caveats regarding the project, especially in relation to Edouard Machery's Concept Eliminativist (CE) project.

First, unlike Machery's CE, CLE doesn't aim to eliminate a commonly used vernacular term. 'Concept learning' and 'concept-learning mechanism' are both scientific terms. Their elimination wouldn't have implications for non-scientific talk about topics related to cognitive development. However, CLE will pose a challenge to the sometimes-loose use of the notions in the philosophical and psychological literature; it will demand that scientists specify which specific mechanism they have in mind when discussing issues of concept learning.

Second, although CLE will employ the style of argumentation that Machery used in his own argument for CE, I want to stress that the two positions are not interchangeable. If the term 'concept' indeed were successfully eliminated, CLE would be a very likely (maybe even trivial) consequence. If, on the other hand, 'concept learning' were successfully eliminated, CE would not necessarily be implied. Even if CL didn't provide a scientifically useful and valuable category, the notion 'concept' could still be sufficiently unified because of other factors that are independent of questions of acquisition.

7.3 The Horizontal Argument for Scientific Eliminativism

The aim of Edouard Machery's book *Doing Without Concepts* is to give an argument for eliminating the use of the term 'concept' in cognitive psychology. To this end, Machery discusses a large amount of psychological research in concepts, and formulates what he calls the Heterogeneity Hypothesis. The main claim of the Heterogeneity Hypothesis is that a very large part of the categories humans use for conceptual thought "are represented by several concepts that belong to kinds that have little in common" (Machery, 2009, p. 60). Machery is however not content with just showing the heterogeneous nature of concepts, but wants to draw a radical conclusion from this:

> The notion of concept ought to be eliminated from the theoretical vocabulary of psychology because it might prevent psychologists from correctly characterizing the nature of the knowledge in long-term memory and its use in cognitive processes. (Machery, 2009, p. 220)

In this section, I will go through Machery's argument for this conclusion, and will thereafter apply it to the case of concept learning in Section 7.4. My own conclusion is that concept learning doesn't fare well when confronted with this argument, and that the consequence from this is the elimination of the notion 'concept learning' from psychological and philosophical theorising.

7.3.1 Types of eliminativist arguments

Machery (2009) discusses two general classes of eliminativist arguments – semantic and scientific eliminativism. I will present the most important types of these classes of argument and show that the second type of scientific eliminativism – horizontal eliminativism – is the suitable one for my current enterprise. Introducing the alternatives to Machery's horizontal argument will also avoid the possible confusion of my project with a Churchland-style semantic eliminativist proposal.

Semantic eliminativism is based on arguments from successes or failures of theories of reference. Machery (2009, p. 224f.) describes the general argumentative structure in roughly the following way:

1. Term 'x', as used in explanations of phenomenon P, is defined to mean 'X'/ refer to X.
2. There is evidence (e.g. from a science pertinent to P) which shows that there are no Xs.
3. So, 'x' doesn't refer to anything.
4. Therefore, the term 'x' should be eliminated from discourse about P.

He is critical of this type of argumentative strategy (e.g. in Mallon et al., 2009; Machery, 2009) because it is dependent on the correctness of its presupposed theory of reference – for instance, when presupposing a causal-historical theory of reference, the eliminativist appeal of the argument disappears (cf. Machery, 2009, p. 225). As long as there is no reason to regard one theory of reference as definitive and true, philosophers shouldn't rely on using reference as a key factor in their treatment of other philosophical questions, Mallon et al. (2009) suggest. For this reason, among others, Machery opts for the second type of eliminativist strategy.

Among scientific eliminativist arguments, Machery identifies two general types, embodied by 'vertical' and 'horizontal' styles of argumentation, which I will to turn to now.

The prefix 'vertical'/'horizontal' is used as an indicator of the argument's direction of movement on the scale of levels of explanation. Stich and Murphy (1999) use it when discussing Paul Griffiths's arguments for eliminating the term 'emotion' in Griffiths (1997).

A vertical argument for scientific eliminativism argues that a term should be eliminated from scientific vocabulary because its use at different levels of explanation is inconsistent (cf. Stich and Murphy, 1999, p. 24). Machery notes that the vertical argument for scientific eliminativism allows for the existence of several natural kinds which afford generalisations on different levels of explanation (cf. Machery, 2009, p. 237). The overarching term that is supposed to cover these natural kinds might however mask important differences between them. An example can be found in Griffith's treatment of 'emotion', which he finds covers the natural kinds "affect programs, higher cognitive states and social constructions" (Stich and Murphy, 1999, p. 14). The vertical elimination step consists in acknowledging the existence of all three subclasses of Emotion, while purging the general term 'emotion' to cover them and giving up the idea that the conjunction of the three kinds forms a cluster of concepts worthwhile to be researched as a single cluster.

An analogy may help to understand the process of vertical elimination. Consider the term 'meteor'. As Jankovic (2006) describes, Aristotle (and a long tradition following him) regarded everything that fell from the sky as a meteor, including the common weather phenomena such as rain and snow. 'Meteor' thus covered several phenomena that modern science studies from completely different angles: astronomers study meteors, meteorites and meteoroids (among other things), but don't look at weather and other atmospheric phenomena. This is the domain of meteorology, today understood as the science of the weather. So, what used to be grouped together under the term 'meteor' is today found to fall within the scope of fundamentally distinct sciences, using very different methods and explaining phenomena on different levels of explanation. Eliminating 'meteor' as a term covering the objects falling from the sky thus has the 'vertical' effect of disentangling levels of explanation that don't actually speak to each other: if the term were still used, meteorologists would mean something with it (e.g. rain, snow, possibly meteors, i.e., stones falling from space) that astronomers don't (who only refer to stones falling from space).

In comparison to this 'vertical' style, the horizontal argument for scientific eliminativism remains on one level of explanation, like the cognitive-psychological level in context of concepts like EMOTION or CONCEPT, in contrast to the neurophysiological level or the sociological level. A simple example of a horizontal scientific elimination, albeit one which hasn't translated into folk usage, comes from the concept JADE. As

Hacking (2007), following on Putnam (1975), describes, the concept JADE came to refer to two kinds of minerals, jadeite and nephrite. They are similar in some respects, but are two completely different minerals:

> Jadeite is a combination of sodium and aluminum. Nephrite is made of calcium, magnesium, and iron. (Putnam, 1975, p. 241)

For the mineralogist, the term 'jade' doesn't play a role since it covers two very different kinds of minerals that haven't come to be classed as belonging to the same type because of scientific reasons but simply because at first they looked very similar to both Chinese and Europeans (cf. Hacking, 2007, p. 270f.). As an instance of scientific eliminativism, the case of jade is instructive: at the chemical level, two substances are very different from each other, yet bundled together by one classificatory term. Upon discovery of the first fact, the classification is given up, i.e., the term 'jade' is eliminated from scientific terminology.

In brief, the structure of the Horizontal Argument for Scientific Eliminativism (HASE), as applied to Concept, is as follows:

1. If a "hypothesized natural kind term fails to pick out a natural kind"(Machery, 2009, p. 246) and there are pragmatic grounds for eliminating it, then the natural kind term should be eliminated.

2. 'Concept' doesn't pick out a natural kind.

3. There are pragmatic reasons against keeping 'concept' as a scientific term.

4. Conclusion (MP 1, 2&3): 'Concept' should be eliminated.

The first premise comes down to whether 'concept' picks out a natural kind. The second one, to be treated further below, is whether there are pragmatic considerations in favour of dropping 'concept' as a term of scientific investigation. For the purpose of investigating the elimination of the term 'concept learning' from cognitive psychology, we should follow Machery and use the horizontal argument for scientific eliminativism. The most important reason against using the vertical argument is that the discussion of concept learning in the preceding chapters has been dealing only with the cognitive psychological level. I haven't engaged with the vertical structure of concept learning, e.g., the potential differences between neurophysiological, cognitive, and social theories of CL. The most important reason for not using the semantic eliminativist argument is that it doesn't rely on the most relevant criteria for evaluating a scientific term's role in a theory. Two of the most notable ones are, first, whether a term denotes a natural kind, and, second, whether there are good pragmatic reasons to keep using a term. Rather, the semantic eliminativist argument relies on the reference of a term to an object. Especially in cases like 'concept-learning mechanism', there is no single physical object that is such a mechanism. Additionally, I agree with the criticism of the semantic eliminativist argument that Machery and colleagues have put forward, which makes the argument unattractive for my purposes.

7.3.2 Natural kinds

On a general understanding of what the sciences do, one can regard their aim as finding out what the world is really like and what things there are in the world. The project thus comes down to exposing the structure of the world and the categories of things that are found in it. Since J.S. Mill's times, these questions have been framed by appeal to *kinds* of things (cf. Hacking, 1991). *Natural kinds* are supposed to be the kinds that are at the bottom of the physical sciences, like the chemical elements in chemistry. In the life sciences, like biology and psychology, the question of where to find the natural kinds has been more difficult (cf. Fodor, 1974), and cases like "Are species natural kinds?" have been controversial since Mill's times. Just to round off the vocabulary of kinds before returning to my investigation, I want to mention the two notions of 'kind' that have typically been contrasted with natural kinds. *Nominal kinds* are the kinds that are united just by having been given a name to group them together: 'things that weigh more than 30 kilograms' is a good example, since it shows the arbitrariness that can come with this kind of grouping. A subclass of nominal kinds are *conventional kinds*, which are groups of things that have been created to cluster together things that have some interesting properties in common (as opposed to the things that weigh more than 30 kg) while not being natural kinds in the strict sense introduced above. An example of this is the kind Herbes De Provence, which has a conventional meaning in cooking while still not picking out a set of herbal plants with a deeper botanical connection between them.

Now, we can put this broad understanding of natural kinds to use. When giving an eliminativist argument of the horizontal type, it has to be established that a theoretical notion should be eliminated from scientific theorising if it fails to refer to a natural kind. To give more substance to this idea, I will discuss Machery's thoughts on natural kinds before presenting his reasons in favour of eliminating the term 'concept' from psychology. Machery proposes two constraints on a suitable theory of natural kinds in psychology:

> The suitable account of natural kinds has to satisfy two properties: it has (1) to be applicable to psychological kinds and (2) to be broad, meaning that many classes have to qualify as natural kinds under this account. (Machery, 2009, p. 231)

Machery proceeds to use these criteria on three theories of natural kinds: the essentialist conception, the nomological conception, and the causal conception. The essentialist conception of natural kinds is based on the idea that natural kinds have *essences*,

> that is, a set of intrinsic, causally explanatory properties that are necessary and jointly sufficient for belonging to the natural kind. (Machery, 2009, p. 231)

One can for example characterise the essence of Gold as being a metal with the atomic number 79.

The nomological conception, on the other hand, stipulates that "natural kind terms feature in laws" (Machery, 2009, p. 232). This implies that there must be at least one law for every natural kind, which provides a generalisation that is uniquely satisfied by this natural kind.

The reason why Machery does not want to use these conceptions is that neither of them has the properties Machery demands for a conception of natural kinds. The essentialist picture of natural kinds wouldn't allow for psychological kinds to be natural kinds, since they don't possess essences (cf. Machery, 2009, p. 231). The nomological account won't work because psychological states are usually not governed by laws, but by ceteris-paribus rules (cf. Machery, 2009, p. 232). Since these two fail for his purposes, Machery endorses Richard Boyd (1991)'s "causal notion of natural kinds" (Machery, 2009, p. 232), which I will also refer to as *the Causal Theory of Natural Kinds*:

> A class C of entities is a natural kind if and only if there is a large set of scientifically relevant properties such that C is the maximal class whose members tend to share these properties because of some causal mechanism. (Machery, 2009, p. 232)

In support of this construal, Machery makes the following clarifying points:

- While 'large' is a vague term, it conveys enough to carry out the classificational project at issue. Besides, 'natural kind' is also a vague term.
- The generalisations mustn't be accidental, but must rely on a common causal mechanism.
- The class about which the generalisation is made has to be maximal – there shouldn't be a larger class for which those generalisations hold.

Boyd especially stresses the importance of "'homeostatic property cluster' kinds" (Boyd, 1989, p. 16) for a scientifically accurate understanding of natural kinds. By this, he means that a main reason for calling something a natural kind is that there are several properties of the kind that tend to appear together because of some causally or historically relevant mechanisms; these mechanisms bring about a relative stability (hence *homeostasis*) of the central properties of the kind. Machery agrees that homeostatic property clusters are important for understanding natural kinds, but he wants to allow for a broader foundation for natural kinds: according to him, the causal mechanism underlying the grouping of a set of entities as a natural kind could be an essence, a function, the homeostasis of properties, or common descent (cf. Machery, 2009, p. 233).

One important point for Machery is that the properties of a given class which are relevant to psychologists often aren't the properties by which we identify this class: when spotting birds, you might recognise the blue tit because of the yellow feathers on its breast. According to Machery's picture, however, the yellow breast wouldn't count as a property by which you delineate the natural kind Blue Tit, since it is not a scientifically relevant property of blue tits. Ornithologists may however find out (or already have found out) about a given causal mechanism which makes it so that a vast majority of this kind of bird has a yellow-feathered breast: the property of having this causal mechanism, e.g., this genetic make-up, can count as a scientifically relevant property, and thus as a factor in classifying Blue Tit as a natural kind.

This point is relevant to Machery's argument for concept eliminativism. To arrive at premise (2), Machery denies that 'concept' picks out a class of entities about which psychologists can make many generalisations, apart from the generalisations used to identify concepts. Identifying concepts as those bodies of knowledge that are stored in long-term memory which are used for higher cognitive functions is one thing – but psychologists want to find out the properties of concepts that are responsible for bringing about these higher cognitive functions, e.g., storing a certain type of body of knowledge in long-term memory. These properties are likely to not be the same for different theories of concepts; forming a stable theory of the concept ELECTRON will require other psychological mechanisms than forming a stable statistical representation of the features of blue jays. What happens is that many scientifically relevant generalisations only fit a subclass of the supposed natural kind, depending on the theoretical allegiances of psychologists: prototype theorists make generalisations about prototypes, exemplar theorists do this for exemplars, and theory theorists do it likewise for theories. For Machery, these subclasses of concepts might well be natural kinds by his definition, even if their covering term, 'concept', doesn't correspond to a natural kind itself.

7.3.3 Pragmatic grounds

Consider a term of a given scientific theory that has been identified as failing to refer to a natural kind. Let's suppose Machery is right and 'concept' does not refer to a natural kind. Machery's argument introduces a further test for 'concept' elimination before declaring the term eliminable – the pragmatic considerations regarding its place in cognitive psychology. In the case of 'concept', one might find several pragmatic features of present theorising that seem beneficial for research in concepts. First, Machery proposes that a term might be useful because it is convenient shorthand for an important description. For example, 'concept', could be shorthand for "'bodies of knowledge used by default in the processes underlying most higher cognitive competences.'" (Machery, 2009, p. 239) Second, one might think it an advantage to theorising to have a more general term under which to group a wide variety of research. So, having the label 'research in concepts' might be beneficial because it groups together work in several disciplines (philosophy, cognitive psychology, and comparative psychology).

Now, all of these pragmatic points might not be decisive in the case of a given scientific term that doesn't cover a natural kind if there are serious pragmatic disadvantages to its inclusion in a theory. Generally speaking, there are several pragmatic reasons to give up a scientific term. I will give more examples in the next section of this chapter, but I want to mention two general kinds of consideration that pertain to a discussion of the pragmatics behind scientific terms.

One pragmatic reason for giving up a term is its lack of explanatory use: if a term fails to be useful in explanation and prediction, then there is a strong incentive to give it up. A second pragmatic reason for elimination is the harmfulness of a term. The kind of harm I take primary would be hindering theoretical progress in a given science; this hindrance can come in a variety of flavours. For Machery, 'concept' is a paradigm example of this: a lot of research energy would be freed up if there were no race for finding and defending the single correct theory of concepts. Accepting concept eliminativism would relieve researchers from competing for the pole position. Instead, they could work on the specifics of their own proposals, and find the single correct kind of prototype theory, exemplar theory, or theory theory.

Another kind of harm, closely connected to the issue just raised, would come from inhibiting the construction of a new taxonomy in a given science because of the adherence to a non-natural-kind term. Giving up 'concept' might open the possibility of "the development of a new classification system that would identify the relevant natural kinds" (Machery, 2009, p. 239).

I will add a few more CL-specific points concerning pragmatics below. I conclude for now that pragmatic considerations can play an important role in shaping our scientific vocabularies.

Now that I have established a framework for examining the possibility of eliminating scientific terms, the next step is to apply it to 'concept learning'.

7.4 The argument for Concept Learning Eliminativism

We have introduced a viable conception of natural kinds and scientific eliminativist argument in the last section. Now, we can make use of them by turning our attention to the main topic of the investigation: the status of the study of concept learning as a research project in cognitive psychology.

The first step in making a case for CLE lies in formulating the structure of HASE with regards to concept learning:

HASE-CL If the term 'concept-learning mechanism' (CLM) doesn't pick out a natural kind and if there are pragmatic reasons against keeping 'concept-learning mechanism' as a term of research in concepts, then 'concept-learning mechanism' should be eliminated from scientific discourse.

To arrive at the eliminativist conclusion, I need to show that the conjunctive antecedent of (HASE-CL) is true. To spell it out, I need to argue for the following two claims:

I The term 'CLM' doesn't pick out a natural kind.

II There are pragmatic advantages to the elimination of 'CLM' as a term in philosophy and psychology.

If (I) and (II) go through, then 'concept-learning mechanism' and by consequence 'concept learning' should be eliminated as scientific terms. Here is the reason why the elimination of the first logically implies the elimination of the second. If there is no unified mechanism for concept learning, then the notion 'concept learning' doesn't play a role in understanding the details of conceptual development, and CLE is the better position with regards to structuring research. In this picture, one can still talk about concept learning in a loose sense, or as a broad functional description of certain cognitive phenomena (just as Machery can still talk about concepts while attempting to eliminate the term from science), but one cannot claim that the term 'concept learning' has any explanatory role in cognitive psychology.

As an analogy, consider the study of memory. There is a clear consensus that the term 'memory' doesn't denote any single cognitive mechanism, but only groups together a diverse set of mechanisms for easy reference. In this sense, 'memory' might still have a role in vernacular discourse ("Moments ago I still knew which subway line to take to MOMA, but when I got distracted, it escaped my memory"), but not in the scientific study of how humans retain and access information.

With this objective established, I will go on and give arguments substantiating claims (I) and (II), respectively.

7.4.1 Against CL as a natural kind

Machery gives several possible ways of establishing that a given term fails to pick out a natural kind. Among those, the pertinent proposal for CL is the case in which

> [t]here are very few generalizations that are true of the K's [instances of the natural kind K], besides the properties that are used to identify the K's. At the same time, many generalizations are true of the members of subclasses of K—K_1,..., K_n. (Machery, 2009, p. 237f.)

The other cases are such that the generalisations are either accidental or are actually generalisations that hold for a superset of K. These are however not relevant for our present concern. With this in mind, we can formulate the argument that CLM is not a natural kind:

1. If something is a natural kind, then there are many scientifically relevant generalisations about a majority of its instances.
2. There are very few scientifically relevant generalisations about the CLMs I investigated so far.
3. Conclusion (MT 1, 2): CLM is not a natural kind.

In the following, I will first show why premise 1 is uncontroversial given the acceptance of the picture of natural kinds on which Machery relies, and which I accept. After this, I will develop the justification for premise 2.

Premise 1: A necessary condition for being a natural kind

As we have seen above (in Section 7.3.2), premise 1 encodes important ingredients of the Causal Theory of Natural Kinds. In fact, it specifies a minimal necessary condition for being a natural kind. These ingredients are the following:

- We need *many* generalisations to justify seeing something as a cluster of properties.
- We need *scientifically relevant* generalisations, which aren't the identifying generalisations.
- Generalisations covering *a majority* of its instances are sufficient, since we only need ceteris paribus generalisations, not laws (which would cover all instances).

At this point, we don't even need to include the second main demand of the Causal Theory of Natural Kinds, namely that there be a causal mechanism thanks to which members of a natural kind are clustered together. In this sense, the condition we need for the argument to go forward can be considered minimal. Thus, if we agree that the Causal Theory of Natural Kinds is adequate for investigating psychological terminology and theorising, then we have premise 1 as an uncontroversial component in our argument.

Premise 2: The scientifically relevant generalisations about the different theories

We now turn to the theories that I discussed in the previous chapters, as they are the best evidence for scientifically relevant differences between CLMs. A plausible expectation for a generalisation about concept-learning mechanisms would be that it has roughly the following form:

CLM Generalisation Ceteris paribus, most CLMs operate on target domain D using the causally active mechanisms $M_1, \dots, M_n$.

One option for choosing a domain D would be to demand that a CLM has to be able to account for all concept learning, thus demanding D = {All kinds of concepts}. This would however be unreasonable given the CLMs we have looked at. For instance, neither SBSMs nor Perceptual Learning (PL) can be expected to explain the learning of abstract concepts and non-perceptual concepts (unless one is convinced by Barsalou's and Prinz's reconstructions of abstract concepts in perceptual terms, see Barsalou (1999); Prinz (2002)). The HF model and the Sellarsian model, qua learning models based on the use of language, can account for abstract and non-perceptual concepts, but not as easily as for purely perceptual concepts, such as auditory concepts or 'Martian Rock' concepts (as in Chapter 4).

To make matters worse, it doesn't look like the CLMs we have discussed rely on a shared set of *defining* mechanisms for concept learning. The only mechanism which all of them need is a perceptual apparatus, which gives the learning mechanism its inputs. But except for the case of PL and the Pask device, this isn't the point at which the explanation of concept learning starts. Rather, it is a set of cognitive biases and inferences which guide the learning process, as in the case of joint-attentional learning and SBSMs. Even here, joint-attentional learning and SBSM-learning come apart as they rely on different kinds of inferences and biases: SBSMs rely on syndrome construction and the essentialist bias, whereas joint-attentional learning primarily relies on social-cognitive mechanisms such as intention-reading.

Just by looking at the explanatorily central mechanisms it becomes evident that these CLMs are diverse in their mechanistic setup. There is not enough coherence between these CLMs to justify treating them as examples of one family of homeostatically clustered causal mechanisms, as the Causal Theory of Natural Kinds would demand. Additionally, when focusing on the subkinds of CL, we can see that there are several interesting generalisations possible – which makes them more interesting to study on their own, detached from the question whether they *are (the only/main kind of)* concept learning. Take Goldstone's PL as an example: on a prudent interpretation, generalisations about perceptual learning mechanisms have a well-defined target domain D_{PL} = {Perceptual Concepts}. Adding the set of mechanisms bringing about PL that we have reviewed in Chapter 4 to the generalisation, we get the following as a starting point for further investigation and further generalisations about Perceptual Learning:

PL Generalisation Ceteris paribus, PL operates on the target domain of perceptual concepts using the causally active mechanisms of unitisation and differentiation.

If we were to rephrase the descriptions of the hitherto-discussed CLMs we could form generalisations of a similar form that apply exclusively to that one CLM. I propose to move on from the analysis of scientific generalisations about CLMs and I regard the following propositions as established.

There are good reasons to assume that all of these CL mechanisms rely on different explanatorily central features, even though their goal is broadly speaking the same. The only generalisation about all the CLMs I have reviewed is that they all share the same function, generally speaking: they all have the function of learning concepts. This, however, is not a scientifically relevant generalisation, but only an identifying generalisation, and I have ruled these out as evidence for something's being a natural kind in Section 7.3.2.

With these main elements in mind, I will conclude that there is a strong basis for the first claim of the CLE argument.

7.4.2 Pragmatic reasons for CLE

Just as with the term 'concept', there may be some pragmatic advantages to keeping the terms 'concept learning' and 'concept-learning mechanism'. I will argue that there are four important pragmatic considerations in favour of CLE that weigh more heavily than the pragmatic considerations in favour of keeping CL and CLM.

A first and major pragmatic point in favour of using the terms CL and CLM in cognitive-psychological research is that they cluster together all the theories and potential mechanisms of learning new concepts. This might be seen as pragmatically advantageous because it enables researchers to effectively compare their proposals with other CLMs that attempt to explain the same learning phenomena.

A second pragmatic advantage of keeping CL is that it serves as shorthand for the variety of mechanisms I have discussed. It might be too cumbersome to refer to a list of CLMs or to a detailed functional description when one just wants to talk about the broad class of concept-learning mechanisms.

Against these reasons, I will pit the following pragmatic reasons to endorse CLE.

Scientific Progress CLE promotes scientific progress in research on concepts.

Scope CLE improves the explanatory success of theories by modifying their scope.

Variability CLE reflects the wide variety and variability of learning environments and learning domains, as opposed to a 'one-size-fits-all' theory.

Explanatory simplicity There is higher explanatory simplicity in individual theories based on subkinds of CL.

In the following sections, I will go on to show that there are good reasons to accept these points. If they all hold, I think that they can trump any pragmatic advantages of keeping CL. I will first present the arguments in short enumerated form and then elaborate on their implications. Where appropriate, I will give concrete examples of advantages for research in concept learning.

Scientific progress

1. Research in a given area can make more progress when it drops ineffectual research efforts.
2. The quest for proposing the main, and uniquely right theory of CL can stall the contender theories' progress in their core domains of research.
3. Without the quest for the uniquely right theory of CL, research in CL can make more progress.
4. When accepting CLE, theorists of CL can drop the question of finding a uniquely right theory of CL altogether.
5. Therefore, CLE promotes scientific progress in research on concepts.

Considerable research effort usually goes into safeguarding one's theory against its competitors and establishing its advantages, for example in explaining a set of empirically tested phenomena. This is based on the assumption that only one theory can be right about the fact of the matter: concept learning can only be one thing, one kind of mechanism, and granted one is convinced to lead research in the 'right' theory of CL, it becomes a necessary part of one's research to establish arguments and evidence that demonstrate the superiority of one's theory. This can take many forms, and many of them might further our knowledge about concept learning, such as pointing out explanatory lacunae, flawed experimental designs, and the like, in competing theories. All of this is just part of good science. If my argument goes through, doing this in order to recommend one's own theory is just binding resources that would be better used in running experiments and writing papers to strengthen one's own research project.

If we accept CLE, the aspect of 'theory safeguarding' falls away, and research efforts can instead be directed towards solidifying evidence for a theory's empirical adequacy and towards specifying the details and foundations of the theory. With CLE, current theories of CL don't have to compete for being the main and uniquely right theory of CL, and scientists can direct their research to more fruitful questions about their own theory. I take this to be one of the main factors that make Concept Learning Eliminativism recommendable.

To illustrate the advantage it brings, let's consider the example of Perceptual Learning. Once we grant the possibility of learning concepts through PL, and also grant that it needn't compete with, for instance, joint-attentional learning, we can go ahead and focus on the question of the specifics of PL: should we follow Landy and Goldstone (2005) in assuming that it operates with unitisation and differentiation, or should we consider additional mechanisms, like those of Bhatt and Quinn (2011)? Following from this, we can devise new tests for the proposed coexistence of unitisation and differentiation; we can use the models to gain data from studying new groups of subjects (infants, young children, older-age persons, brain-lesion patients, ...), and possibly many more things. All of this isn't needed to safeguard PL against other competitors, but to strengthen the research project itself, and to increase its explanatory range.

I conclude from this that CLE promotes scientific progress.

Scope

1. Every theory of CL I have reviewed has difficulties explaining the learning of some concepts.
2. Limiting the explanatory scope of these theories increases their explanatory success.
3. CLE is compatible with, and encourages, scope adjustment in theories of CL.
4. Therefore, CLE improves the explanatory success of theories by modifying their scope.

The worry behind this argument has already been raised in section 7.4.1 above, when I observed that there is no unique domain of target concepts that every CLM can aim to learn. There are several demands on a complete theory of concept learning that the available contenders have difficulty fulfilling.

The scope of target concepts isn't the only dimension worth mentioning here. Quite likely there are ontogenetic differences between CLMs coming online. Just to give one example, joint-attentional learning depends on the development of gaze-following and pointing, and this won't come online until around 9-12 months of age. It would therefore be wrong to assume that joint-attentional concept learning can explain CL across the whole lifespan. By comparison, Perceptual Learning might play an important role from prelinguistic infancy onwards. The lesson I want to draw from these examples is this: pragmatically, it is preferable to relieve the individual CLMs of the burden to explain phenomena they aren't fit to explain by accepting CLE. The plurality of CLMs is better suited to answering the range of demands on a theory of CL. Eliminating the notions CL and CLM ensures there is no doubt about the scope and application of each individual mechanism.

Variability

1. The circumstances and environments of CL are diverse and variable.
2. A unified theory of CL has difficulties accommodating this variability.

3. CLE complements variability in CL.
4. Therefore, CLE has a pragmatic advantage to a unified theory of CL.

First, this isn't the same point as the one expressed in the (Scope) argument: while (Scope) concerns the targets of concept learning – the kinds of concepts whose learning a CLM can explain – (Variability) concerns the conditions under which any kind of concept can be learnt. Consider an example: you can learn the concept BLUE JAY by observing the birds in the wild, but you can also learn it by reading about them in a book. You can learn the concept VELOCITY during driving lessons, or you can learn it in a physics class. The point of the (Variability) argument is that any CLM that we have considered would have difficulty in explaining at least some circumstances of successful concept learning. It seems likely that a diverse set of CLMs has the pragmatic advantage of explaining different learning situations in different ways. Joint-attentional learning cannot explain the lonesome birdwatcher's progress in telling apart ravens, mockingbirds, and blue jays. At the same time, perceptual learning wouldn't be the optimal model to explain the learning of the concept CREATIVITY in an art-class discussion. Any one CLM is insufficient to cover the learning environments and circumstances in which we think concepts are learnt. By dividing them among the available CLMs, we take away a further reason to regard CL as unified. Pluralism about concept learning – CLP, as introduced at the end of Chapter 6 – is not tenable in the face of (Variability) and (Scope), because there are too few commonalities between the different clusters of learning mechanisms, their environments, and also their scope.[2] This point feeds back into the natural-kinds argument from Section 7.4.1, and also reveals a pragmatic advantage of CLE. A theory that is more suited to explain an important but isolated part of a set of phenomena can do so more effectively than one that has to be contorted to do so by adding caveats or exceptions. This leads directly to the next point.

Explanatory simplicity

1. A unified theory of CL needs several caveats and conceptual divisions to function.
2. *Ceteris paribus*, simpler explanations are preferable to complex explanations.
3. A diverse set of theories of CL, which is a result of CLE, gives simpler explanations since it works without the complexities of a unified theory.
4. Therefore, CLE affords higher explanatory simplicity for theories of CL.

There is a higher explanatory value in individual theories based on varieties of CL – one general theory (i.e. a theory based on the kind 'CL') that attempts to accommodate the plurality of CLMs would have to come with a lot of caveats and subdivisions. Pragmatically, this makes it more difficult to work with the theory. Simpler, non-unified theories explain the phenomena in a more straightforward way while still jointly

[2] I will say more about CLP in Section 7.5.

covering the range of explananda. I take this sort of explanatory simplicity to bring added explanatory value to CLE.

The bottom line

In conclusion, I submit that the pragmatic reasons I have given for CLE outweigh the pragmatic reasons for keeping CL. Especially the issues of scientific progress and explanatory simplicity speak strongly for CLE, with scope and variability also counting against the idea of a unified theory of CL. A sophisticated version of Concept Learning Pluralism (CLP) might be able to accommodate the latter two, but would still face the problems of the former two. For now, I want to postpone the detailed discussion of CLE vs. CLP with specific regard to the issues of this section, but I will raise some concerns about CLP in the following section on the problems with CLE.

7.5 Problems with CLE

Having offered an argument for the eliminativist hypothesis about concept learning, I recognise that there are a number of possible objections. Some of these objections were already directed at Machery's CE in the comments section of Machery (2010), so I will use this as a starting point. Here is a short summary of the objections that I am going to address:

1. We should just be happy to be CL Pluralists instead of eliminativists.
2. We should hold that the HASE doesn't work for CL mechanisms, since mechanisms can't be natural kinds.
3. We should worry about the overgeneralisation that CLE allows for: if it is that simple to eliminate a scientific term, wouldn't a thorough look at scientific practice, or at the philosophical use of scientific theories, radically reduce the vocabulary scientists can use, to the detriment of theory construction?
4. We should say that we're not far along enough to make a verdict: maybe it will turn out that there is one primary CL mechanism that is at the bottom/centre of all concept learning.

7.5.1 Concept Learning Pluralism

Suppose that you accept the point that there are several different kinds of concept learning. Suppose you are convinced that these kinds of CL don't have a lot in common, as I have argued in the previous section. Still, you might not want to accept CLE. Why isn't it enough to be a Concept Learning Pluralist? I take the following to be the main reason for being a Pluralist.

Hampton (2010) argues in the case of concepts that the notion 'concept' is necessary to structure psychological research, and that the link between the subkinds of 'concept' (prototypes, exemplars, and theories) wouldn't be clear without the notion 'concept' uniting them:

> First, there is the obvious point that the [...] representations [...] of dog [sic] all refer to the same class – they are broadly coreferential [...]. What makes them co-referential is the fact that they represent the same concept. (Hampton, 2010, p. 212)

An analogous point might be made about concept learning: without the notion 'concept learning', it wouldn't be clear what the different subkinds of CL have in common. You can be a Pluralist about concept learning only if you acknowledge that the different kinds of CLMs all broadly speaking have the same aim, or are related to the same general function.

I think that this point doesn't weaken CLE since it doesn't translate into a proper challenge: being united under a natural kind is not essential for explaining different kinds of concept learning. Let me illustrate with Hampton's example of the concept DOG. Suppose that you can learn DOG by several CLMs; you might even learn different DOG-concepts because of this. This obviously has consequences for the theory of concepts one endorses, and it might explain phenomena such as inconsistencies in judgments about dogs, but it doesn't lead to any need to unite the CLMs under one common heading.[3]

Put bluntly, I think that the CL eliminativist gets all the benefits that the CL Pluralist has, but doesn't have to accept the shortcomings of structuring the research landscape with CL as a key ingredient. Additionally, the Pluralist doesn't get all the pragmatic benefits I have identified above (although she might get some, e.g. (Scope)).

7.5.2 Mechanisms as natural kinds

A second broad strategy to attack CLE regards the status of mechanisms as natural kinds. The most general claim in this direction is that mechanisms needn't or even shouldn't be seen as natural kinds, and that this damages the eliminativist position. Again, there are different flavours of this suggestion:

- Lalumera (2010) argues, in the case of concepts, that 'concept' doesn't need to be a natural kind, but still shouldn't be eliminated because it is a functional kind.
- Mechanisms generally can't be natural kinds because they don't have essences.

[3] On an aside, I think Hampton's point also doesn't seriously challenge Machery's position, since he can reply to it that it's not the concept DOG that unites prototypes, exemplars and theories of dogs, but it's the referents, and that reference and categorisation can come apart (cf. Machery, 2010, p. 235).

As with the previous problem, I don't think that this set of worries is a decisive point against CLE. In response to Lalumera's suggestion, I think it is fair to follow Machery's own response (cf. Machery, 2010, p. 238). While 'concept learning' may be a functional kind, this doesn't settle the desirability of using the term within a psychological framework: if a functional kind doesn't track a natural kind or bring some pragmatic advantages for the science in question, then its functional nature doesn't weigh heavily enough against its elimination.

In response to the second worry, I think that it is fair to refer back to Machery's point about the applicability of the essentialist view of natural kinds to psychological kinds in general. It is far from clear whether psychological kinds *can* have essences in the first place. Besides this, the essentialist view can be challenged because of its failure to capture some supposedly genuine natural kinds (cf. Strevens, 2000). In my own interpretation, the notion 'essence' is too general to track what it means to express, and more specific descriptive devices are better suited to do their work, such as genes in biological categorisation, or chemical structure in non-biological natural categorisation. Within this alternative framework, there is room for psychological natural kinds, since no one would expect that psychological kinds have, for example, a determined genetic makeup (although there might be other, more naturalistic foundations for psychological kinds, which I cannot discuss here).

7.5.3 Overgeneralisation

If it is this simple to eliminate a scientific term, wouldn't a thorough look at scientific practice, or philosophical use of scientific theories, radically reduce the vocabulary scientists can use, to the detriment of theory construction? After all, these theories use lots of notions that most likely won't be natural kinds, such as "representation, module, algorithm, and nutrient." (Machery, 2010, p. 238) This worry is levelled against Machery's CE, for instance by Gonnerman and Weinberg (2010) and Khemlani and Goodwin (2010).

In response to this, I want to stress a few points. First, I take it that the empirical sciences are in the first instance not vocabulary-driven but data-driven in that scientists have the ability to start theorising about a given topic by researching the phenomena. So, even when acknowledging the theory-ladenness of observation in science, the bulk of theory-construction comes after gathering observations. Eliminating notions that don't play a role that goes along with a theory's observations might limit the kinds of theories we can construct, but if the criteria of elimination are good, then they limit theory-construction for the better. I have given arguments showing that this is indeed the case for CL, so I take this to already be a sufficient reason to resist the overgeneralisation worry.

Second, I take Machery's own response to hold equally well against the worry (cf. Machery, 2010, p. 238f.). He repeats that natural-kind status is not the only criterion against eliminating a notion, but that pragmatic considerations weigh just as heavily. Additionally, for many theoretical notions that are not natural kinds, there are no attempts to discover scientifically interesting generalisations about them. I take Machery to mean that these notions would fly under the radar of eliminativists on the look for new targets, and that these notions could just go on to be employed.

To summarise, I think that this problem doesn't harm the case for CLE. It resembles some typical slippery slope arguments which turn out to be fallacious or not to apply to a situation because the slope isn't as slippery as one might expect, i.e., there are good criteria where to stop eliminating.

7.5.4 A primary kind of CLM

Maybe it is too early to abandon the search for a unifying, or unique, view of concept learning. It might turn out that there is one general mechanism at the bottom of our concept acquisition, and this mechanism will be all that we need to understand concept learning.

One way to answer this worry is to look at the prospects of the presently available theories of concept learning. Do any of these theories hold the promise of being the primary kind of CLM, provided that we find out more about them? As I have laid out in previous chapters, and in the comparison of the different kinds of CLMs in Section 7.4.1, the prospects don't look good. Even the most promising theories, such as the SBSM, cannot account for the acquisition of all kinds of concepts. Perceptual Learning is similarly limited, being unable to explain a large chunk of concept learning. My proposal is to eliminate the notions 'concept learning' and 'CLM', at least until such a genuinely unifying CLM can be found. Jerry Fodor would perhaps suggest that we shouldn't hold our breath.

My present conclusion is that the objections against CLE don't cut the ice. So, my conclusion from this investigation is that CLE is a viable position. In the final section, I will rehearse the consequences that CLE brings to research in concept learning.

7.6 The consequences for concept learning

I want to emphasise the advantageous consequences of accepting CLE. These consequences overlap with the reasons given in Section 4.2. In a sense, pragmatic reasons to adopt a certain type of theoretical outlook are bound to have the kinds of beneficial consequences for the domain of study in question. I would like to reiterate this set of consequences, and end with an observation regarding Jerry Fodor's stance on concept learning.

Refinement of individual theories

Since there is no need to compete for the position of *the* single concept-learning mechanism, psychologists can research the specifics of the models that come to be accepted kinds of concept-learning mechanisms, for instance the question of which account of Perceptual Learning should be favoured.

Ontogenetic conceptual development

Especially in developmental psychology, there is a fierce competition between different kinds of concept-learning mechanism, as the investigation in Chapter 3 has shown. As a reminder, that chapter discussed two important theories of concept learning in developmental psychology – Jean Mandler's and Susan Carey's, who represent an 'Empiricist' and a 'nativist' position, respectively. While I had concluded that Mandler's theory doesn't overcome Fodor's Challenge, and that, in its present form, it shouldn't be endorsed as an account of CL, I have also hinted at possible ways to redeem Perceptual Meaning Analysis (PMA). One suggestion was to give up the strict distinction between perceptual and conceptual representation, at least in the manner that Mandler draws it. Assuming that an 'update version' of PMA is possible, we gain the following new perspective on concept learning in early ontogeny.

If CLE is correct, the two opposing theories can be seen as complementary, and they can be integrated into a single ontogenetic timeline with a division of labour between Mandler's PMA and Carey's Quinian Bootstrapping. Furthermore, other developmental theories, such as Tomasello's Social-Pragmatic Theory can add to the mosaic of ways of learning concepts. This doesn't mean that there won't be disagreement between the proponents of the different mechanisms, but it is more likely that disagreement won't be about the questions related to concept learning, but to other aspects of the human mind.

Theories of concept acquisition

If we eliminate the term 'concept learning' from the cognitive sciences and say that there is no such thing as concept learning, then 'concept acquisition' becomes the term under which we should class all the concept-learning mechanisms that I have discussed in this thesis. While this may seem like an undesirable consequence because it breaks down the boundary between instances and mechanisms of brute-causal acquisition and rational acquisition/learning, I want to propose that this is actually an advantage.

First off, I think that 'concept acquisition' is a lot less likely to be considered a natural kind term, due to its very broad scope. For the purpose of classifying instances and mechanisms of concept acquisition, the term should be regarded as covering a functional kind – what unites all mechanisms of concept acquisition is that they are doing what their name implies.

Furthermore, we can employ Margolis and Laurence (2011)'s criteria for concept learning, which I have discussed in Chapter 6, to bring order to the range of concept-acquisition mechanisms. With CLE, the distinction between 'brute-causal' acquisition and 'rational-causal' learning is removed and we have a more flexible tool to evaluate the success and versatility of concept-acquisition mechanisms at hand. Additionally, we can use further criteria, such as the scope or the applicable cases of a concept-acquisition mechanism, to determine how well a particular mechanism is suited to explain the acquisition of a given concept. The move away from 'concept learning' thus will not lead to a confounded field of research, but rather to a clearly differentiated set of acquisition mechanisms.

If we conceive of the mechanisms of concept acquisition in this sense, then new ways of thinking become available: for example, swallowing a pill to 'learn' a set of Latin grammar laws can be regarded as a better, or more rational, way of acquiring concepts than other pill-ingesting methods. What counts as good cases of concept learning depends on more mobile boundaries, and on pragmatic interests, and no longer on a binary division between the brute-causal and the rational-causal.

The consequences for Fodor

The prospect of CLE echoes Fodor's conclusion that there is no such thing as concept learning – there is no single unified class of concept-learning mechanisms, and no reason to pluralise (i.e. to replace CLE with CL Pluralism). As a reminder, the conclusion of Fodor's paradox is that all concepts are innate or acquired in some non-rational way. Fodor's conclusion is reached by showing that the path to concept learning is blocked by a paradox that makes the concept-learning mechanism dependent upon the concepts it is supposed to be learning because they are required for forming hypotheses about the concept.

Surprisingly, Fodor's verdict that there is no such thing as concept learning comes out as true on my view, albeit for a *completely* different reason. Instead of establishing CLE by Fodor's paradox, I have given reasons against accepting Fodor's argument, and I have developed an empirically based argument that relies on considerations from the philosophy of science.

This wouldn't make the argument any more acceptable for Fodor. My argument is premised on the refutation of his argument against the possibility of learning concepts: recall that the bulk of this investigation has been dedicated to showing that Fodor's empirical premise about the HF model as the only available CLM is false. Only once this premise was established could we go on and discuss the differences between CLMs.

Conclusion

To conclude this investigation, I want to emphasise its main results and raise some questions for future research. I started by appraising the role of concepts and concept learning in the philosophy of mind and in psychology, and found that concepts are central in our understanding of human cognition. Furthermore, the power to explain concept learning was found to be of importance for theories of concepts. Thus, I have devoted my investigation to this particular aspect of human cognition.

In order to provide a clear understanding of the subject matter, I set about to explore one main challenge to theories of concept learning, which I found in Jerry Fodor's Learning Paradox. After providing an analysis of Fodor's paradox, I discussed his own answers and found them wanting – his appeal to non-rational concept acquisition faced serious problems. Instead of resigning and accepting that neither concept learning nor other kinds of concept acquisition were possible, I chose to investigate philosophical and psychological models that have been specifically proposed for countering Fodor.

There are two major themes to the concept-learning mechanisms I investigated. One group tries to explain concept learning by giving a perceptually-based model that can turn perceptual information into concepts, loosely speaking. The three models of this kind which I discussed are Jean Mandler's Perceptual Meaning Analysis, Robert Goldstone's Perceptual Learning approach, and Gordon Pask's electro-chemical learning device. I found strong reasons to reject Mandler's model until there is further experimental evidence supporting some of her assumptions, and until she can give a stronger explanation of concept learning that cannot be reinterpreted as an instance of Fodor's hypothesis-formation-and-testing (HF) model, which succumbs to Fodor's paradox. At first, I was also hesitant to regard Perceptual Learning and the Pask device as instances of concept learning, but I offered a set of modifications to the former, which could help overcome Fodor's challenge. Furthermore, I have argued that the Pask device at least partially fulfils Margolis and Laurence's criteria for concept learning, which makes it an attractive model for specific cases of concept learning.

The other group relied on a set of psychological biases and cognitive constraints for their explanations of concept learning. The mechanisms and constraints in question were of a wide variety, ranging from an essentialist bias to a proclivity towards theory construction to social-cognitive biases. The models in this vein were Susan Carey's Quinian Bootstrapping, Wilfrid Sellars's pattern-governed learning, Michael Tomasello's Social-Pragmatic Theory, and Eric Margolis's and Stephen Laurence's Syndrome-Based Sustaining Mechanism (SBSM) model. In this set, I found several promising contenders against Fodor. Especially Quinian Bootstrapping, the Social-Pragmatic Theory, and the SBSM model stood out. Sellars's model was marred by its failure to allow for the existence of prelinguistic concepts. In my investigation, I have given good evidence for concept-use in prelinguistic infancy, so this point proved to be a severe problem for Sellars, as well as for philosophers like Gauker who conceive of concepts as words.

With such a large group of working concept-learning mechanisms assembled, I consider Fodor's challenge to be successfully overcome. This is the first main result of my work. But the success of these mechanisms, and the great promise of further progress through future research, necessitates a profound rethinking of the notion of concept learning. As we have seen in the last chapter, there are good reasons to abandon the notions 'concept-learning mechanism' and 'concept learning' as elements of scientific discourse. I have argued that 'concept learning' is neither a natural kind, nor are there pragmatic reasons to keep it as a scientific term. This state of affairs calls for the elimination of the term, and of 'concept-learning mechanism' as well. In my argument, I have used the structure of Edouard Machery's argument for eliminating 'concept' from cognitive-psychological and philosophical discourse. I was able to reply to several points of criticism that were originally put forward against Machery, but also applied to the proposal to eliminate 'concept learning'. With this result, we arrive at an even stronger position with regard to Fodor's paradox – not only have we shown that there are several concept-learning mechanisms that can overcome it, but we have also found good reasons to reject his construal of what 'concept learning' means to begin with.

Let me finally turn to future research that stems from the main results of my thesis. There are several ways in which the elimination of 'concept learning' opens up a new kind of engagement with mechanisms that might be regarded as competitors. One area of research that I am particularly interested in is the interplay between different concept-learning mechanisms in early ontogeny. I think that the intersection between social-cognitive learning mechanisms and theory-based learning mechanisms is a particularly exciting field. As we have seen in Chapters 3 and 5, there are successful models in both fields, particularly Tomasello's Social-Pragmatic Theory and Carey's Quinian Bootstrapping. The first focuses on the role of social cognition in learning, while the second emphasises the importance of theory construction. I think that both meet in the realm of normative constraints on concept learning. Research in social cognition has uncovered young children's early understanding of rules within games, and other normative factors of everyday life that children grasp from around the age of 18 months. This brings up a series of interesting questions: do children learn theories of norms, or does norm understanding emerge from emotional states or external cues? Do normative structures play a role in children's categorisation behaviour and knowledge structures?

These questions have yet to be explored with regards to their philosophical implications and their possible experimental implementations. I assume that the answers we could get from following this line of research would deepen our understanding of infant cognition significantly.

Having taken a look back and a look at potential future projects, I conclude that research in concept learning is far from over, even though it should no longer be seen as one unified research programme.

Bibliography

Adams, F. and Aizawa, K. (1997). Fodor's asymmetric causal dependency theory and proximal projections. *The Southern Journal of Philosophy*, 35(4):433437.

Allen, C. (2009). Teleological notions in biology. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Winter 2009 edition.

Augustine of Hippo (397/1961). *Confessions*. Penguin books, London.

Baillargeon, R. and DeVos, J. (1991). Object permanence in young infants: Further evidence. *Child Development*, 62(6):1227–1246.

Barlow, H. (2001). Feature detectors. In Wilson, R. A. and Keil, F. C., editors, *The MIT encyclopedia of the cognitive sciences*. MIT Press.

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22:577–660.

Barsalou, L. W. (2005). Continuity of the conceptual system across species. *Trends in Cognitive Sciences*, 9(7):309–311.

Bechtel, W. (2009). Looking down, around, and up: Mechanistic explanation in psychology. *Philosophical Psychology*, 22:543–564.

Bechtel, W. and Abrahamsen, A. (2010). Dynamic mechanistic explanation: Computational modeling of circadian rhythms as an exemplar for cognitive science. *Studies in History and Philosophy of Science*, 41:321–333.

Bechtel, W. and Wright, C. D. (2009). What is psychological explanation? In Calvo, P. and Symons, J., editors, *Routledge companion to philosophy of psychology*, pages 113–130. Routledge.

Bhatt, R. S. and Quinn, P. C. (2011). How does learning impact development in infancy? The case of perceptual organization. *Infancy*, 16(1):2–38.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94(2):115–117.

Block, N. (1987). Advertisement for a semantics for psychology. *Midwest Studies In Philosophy*, 10(1):615–678.

Block, N. (1998). Conceptual role semantics. In Craig, E., editor, *The Routledge Encylopedia of Philosophy*. Routledge.

Boyd, R. (1989). What realism implies and what it does not. *Dialectica*, 43(1-2):529.

Boyd, R. (1991). Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies*, 61:127–148.

Brooks, R. and Meltzoff, A. N. (2005). The development of gaze following and its relation to language. *Developmental Science*, 8(6):535–543.

Burns, B. and Shepp, B. E. (1988). Dimensional interactions and the structure of psychological space: The representation of hue, saturation, and brightness. *Perception & Psychophysics*, 43(5):494–507.

Burnyeat, M. F. (1987). The inaugural address: Wittgenstein and Augustine De Magistro. *Proceedings of the Aristotelian Society. Supplementary Volume*, 61:1–24.

Carey, S. (1985). *Conceptual change in childhood*. MIT Press, Cambridge, MA.

Carey, S. (2002). The origin of concepts: Continuing the conversation. In Stein, N. L., Bauer, P. J., and Rabinowitz, M., editors, *Representation, Memory, and Development: Essays in Honor of Jean Mandler*, pages 43–52. Psychology Press.

Carey, S. (2009). *The origin of concepts*. Oxford University Press, Oxford.

Cariani, P. (1993). To evolve an ear: Epistemological implications of Gordon Pask's electrochemical devices. *Systems Research*, 10:19–33.

Churchland, P. S. (1978). Fodor on language learning. *Synthese*, 38:149–159.

Cowie, F. (1998). Mad dog nativism. *The British Journal for the Philosophy of Science*, 49(2):227 –252.

Craver, C. F. (2006). When mechanistic models explain. *Synthese*, 153:355–376.

Craver, C. F. (2009). Mechanisms and natural kinds. *Philosophical Psychology*, 22:575–594.

Cummins, R. (1983). *Psychological explanation*. MIT Press, Cambridge, MA.

deVries, W. A. and Triplett, T. (2000). *Knowledge, mind, and the given. Reading Wilfrid Sellars's "Empiricism and the philosophy of mind"*. Hackett Publishing, Indianapolis.

Dretske, F. (1993). Misrepresentation. In Goldman, A. I., editor, *Readings in philosophy and cognitive science*, pages 297–314. MIT Press.

Eilan, N., Hoerl, C., McCormack, T., and Roessler, J. (2005). *Joint attention: Communication and other minds : Issues in philosophy and psychology.* Oxford University Press, Oxford.

Erneling, C. (1993). Why first language learning is not second language learning – Wittgenstein's rejection of St. Augustine's conception of learning. *Interchange*, 24(4):341–351.

Evans, G. (1982). *The varieties of reference.* Oxford University Press, Oxford.

Farroni, T., Csibra, G., Simion, F., and Johnson, M. H. (2002). Eye contact detection in humans from birth. *Proceedings of the National Academy of Sciences*, 99(14):9602–9605.

Feest, U. (2003). Functional analysis and the autonomy of psychology. *Philosophy of Science*, 70(5):pp. 937–948.

Fodor, J. (1974). Special sciences. *Synthese*, 2:97–115.

Fodor, J. A. (1975). *The language of thought.* Thomas Crowell Publishing, New York.

Fodor, J. A. (1981). The present status of the innateness controversy. In *RePresentations: Philosophical essays on the foundations of cognitive science*, pages 257–316. MIT Press, Cambridge, MA.

Fodor, J. A. (1987). *Psychosemantics. The problem of meaning in the philosophy of mind.* MIT Press, Cambridge, MA.

Fodor, J. A. (1990). *A theory of content and other essays.* MIT Press, Cambridge, Mass.

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong.* Oxford University Press, Oxford.

Fodor, J. A. (2008). *Lot 2: The language of thought revisited.* Oxford University Press, Oxford.

Fodor, J. A. and Lepore, E. (1992). *Holism. A shopper's guide.* Oxford University Press, Oxford.

Fodor, J. A. and Pylyshyn, Z. W. (1981). How direct is visual perception? Some reflections on Gibson's "ecological approach". *Cognition*, 9(2):139–196.

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Gauker, C. (1998). Building block dilemmas. *Behavioral and Brain Sciences*, 21(01):26–27.

Gauker, C. (2005). On the evidence for prelinguistic concepts. *Theoria (Spain)*, 20(3):287–297.

Gauker, C. (2011). *Words and images. An essay on the origin of ideas.* Oxford University Press, Oxford.

Geach, P. T. (1957). *Mental acts: Their contents and their objects.* Routledge Kegan and Paul, London.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought.* Oxford University Press, Oxford.

Gelman, S. A. (2004). Psychological essentialism in children. *Trends in Cognitive Sciences*, 8(9):404–409.

Gerganov, A., Grinberg, M., Quinn, P. C., and Goldstone, R. L. (2007). Simulating conceptually-guided perceptual learning. In *Proceedings of the twenty-ninth annual conference of the cognitive science society*, pages 287–292. Cognitive Science Society.

Gergely, G. and Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in cognitive sciences*, 7:287–292.

Gergely, G. and Csibra, G. (2005). The social construction of the cultural mind: Imitative learning as a mechanism of human pedagogy. *Interaction Studies*, 6:463–481.

Gibson, E. J. (1963). Perceptual learning. *Annual Review of Psychology*, 14(1):29–56.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, 49(1):585–612.

Goldstone, R. L. (2003). Learning to perceive while perceiving to learn. In Kimchi, R., Behrmann, M., and Olson, C., editors, *Perceptual Organization in Vision: Behavioral and Neural Perspectives*, pages 233–278. Lawrence Erlbaum Associates.

Goldstone, R. L. and Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition*, 65(2-3):231–262.

Goldstone, R. L., Braithwaite, D. W., and Byrge, L. A. (2012). Perceptual learning. In Seel, N. M., editor, *Encyclopedia of the Sciences of Learning*, pages 2580–2583. Springer US.

Goldstone, R. L. and Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(1):69–78.

Goldstone, R. L. and Landy, D. (2010). Domain-creating constraints. *Cognitive Science*, 34:1357–1377.

Goldstone, R. L., Son, J. Y., and Byrge, L. (2011). Early perceptual learning. *Infancy*, 16(1):45–51.

Gonnerman, C. and Weinberg, J. (2010). Two uneliminated uses for "concepts": Hybrids and guides for inquiry. *Behavioral and Brain Sciences*, 33(2-3):211–212.

Gopnik, A. M. and Meltzoff, A. N. (1998). *Words, thoughts, and theories.* MIT Press, Cambridge, MA.

Griffiths, P. E. (1997). *What emotions really are: The problem of psychological categories.* University of Chicago Press, Chicago.

Hacking, I. (1991). A tradition of natural kinds. *Philosophical Studies*, 61:109–126.

Hacking, I. (2007). The contingencies of ambiguity. *Analysis*, 67:269–277.

Hampton, J. (2010). Concept talk cannot be avoided. *Behavioral and Brain Sciences*, 33(2-3):212–213.

Hepach, R., Vaish, A., and Tomasello, M. (2012). Young children are intrinsically motivated to see others helped. *Psychological Science*, 23(9):967–972.

Hobson, P. (2002). *The cradle of thought: Exploring the origins of thinking.* Pan Macmillan, London.

Jankovic, V. (2006). The end of classical meteorology, c. 1800. In McCall, G., Bowden, A., and Howarth, R., editors, *The History of Meteoritics and Key Meteorite Collections: Fireballs, Falls and Finds.*, volume 256, pages 91–99. Geological Society, London.

Johnson, M. L. (1987). *The body in the mind.* University of Chicago Press, Chicago.

Johnson, M. L. (1995). Incarnate mind. *Minds and machines*, 5:533–545.

Keeler, M. (2012). Is Quinian bootstrapping cognitively explicable? Manuscript.

Keil, F. C. (1989). *Concepts, kinds and cognitive development.* MIT Press, Cambridge, MA.

Khemlani, S. and Goodwin, G. (2010). The function and representation of concepts. *Behavioral and Brain Sciences*, 33(2-3):216–217.

Kuhn, T. S. (1962). *The structure of scientific revolutions.* University of Chicago Press, Chicago.

Kukla, A. (1994). Non-empirical theoretical virtues and the argument from underdetermination. *Erkenntnis*, 41:157–170.

Lakoff, G. (1987). *Women, fire, and dangerous things: What our categories reveal about the mind.* University of Chicago Press, Chicago.

Lalumera, E. (2010). Concepts are a functional kind. *Behavioral and Brain Sciences*, 33(2-3):217–218.

Landy, D. and Goldstone, R. L. (2005). How we learn about things we don't already understand. *Journal of Experimental & Theoretical Artificial Intelligence*, 17(4):343.

Laurence, S. and Margolis, E. (1999). Concepts and cognitive science. In Margolis, E. and Laurence, S., editors, *Concepts*, pages 3–81. MIT Press.

Laurence, S. and Margolis, E. (2002). Radical concept nativism. *Cognition*, 86:25–55.

Laurence, S. and Margolis, E. (2012). The scope of the conceptual. In Margolis, E., Samuels, R., and Stich, S., editors, *The Oxford Handbook of Philosophy and Cognitive Science*, pages 291–317. Oxford University Press, Oxford.

Le Corre, M. and Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, 102:395–438.

Leibniz, G. (1765/1996). *New essays on human understanding.* Cambridge University Press, Cambridge.

Lipton, P. (2000). Inference to the best explanation. In Newton-Smith, W. H., editor, *A companion to the philosophy of science*, pages 184–193. Blackwell.

Lipton, P. (2004). *Inference to the best explanation.* Routledge, London.

Liszkowski, U. (2008). Before L1. A differentiated perspective on infant gestures. *Gesture*, 8(2):180–196.

Locke, J. (1690/2008). *An essay concerning human understanding.* Oxford University Press, Oxford.

Machamer, P., Darden, L., and Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, 67:1–25.

Machery, E. (2009). *Doing without concepts.* Oxford University Press, Oxford.

Machery, E. (2010). Precis of Doing without concepts. *Behavioral and Brain Sciences*, 33(2-3):195–206.

Mallon, R., Machery, E., Nichols, S., and Stich, S. (2009). Against arguments from reference. *Philosophy and Phenomenological Research*, 79(2):332–356.

Malt, B. C. (1994). Water is not H2O. *Cognitive Psychology*, 27(1):41–70.

Malt, B. C. and Johnson, E. C. (1992). Do artifact concepts have cores? *Journal of Memory and Language*, 31(2):195–217.

Mandler, J. M. (2004). *The foundations of mind: Origins of conceptual thought.* Oxford University Press, Oxford.

Mandler, J. M. (2008). On the birth and growth of concepts. *Philosophical Psychology*, 21(2):207–230.

Mandler, J. M. (2010). The spatial foundations of the conceptual system. *Language and Cognition*, 2(1):21–44.

Mandler, J. M. and McDonough, L. (1995). Long-term recall of event sequences in infancy. *Journal of Experimental Child Psychology*, 59(3):457–474.

Mandler, J. M. and McDonough, L. (1996). Drinking and driving don't mix: Inductive generalization in infancy. *Cognition*, 59(3):307–335.

Margolis, E. (1998). How to acquire a concept. *Mind and Language*, 13(3):347–369.

Margolis, E. and Laurence, S. (2011). Learning matters: The role of learning in concept acquisition. *Mind and Language*, 26(5):507–539.

Margolis, E. and Laurence, S. (2013). In defense of nativism. *Philosophical Studies*, 165(2):693–718.

Marshall, P. and Meltzoff, A. N. (2011). Neural mirroring systems: Exploring the EEG mu rhythm in human infancy. *Developmental Cognitive Neuroscience*, 1:110–123.

Medin, D. L. and Ortony, A. (1989). Psychological essentialism. In Vosniadou, S. and Ortony, A., editors, *Similarity and analogical reasoning*, pages 179–195. Cambridge University Press, Cambridge.

Mervis, C. B. and Crisafi, M. A. (1982). Order of acquisition of subordinate-, basic-, and superordinate-level categories. *Child Development*, 53(1):258–266.

Millikan, R. G. (1989). In defense of proper functions. *Philosophy of Science*, 56:288–302.

Millikan, R. G. (1991). *White queen psychology and other essays for Alice.* MIT Press, Cambridge, MA.

Murphy, G. L. (2002). *The big book of concepts.* MIT Press, Cambridge, MA.

Murphy, G. L. and Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92:289–316.

Nersessian, N. (1992). How do scientists think? Capturing the dynamics of conceptual change in science. In Giere, R., editor, *Cognitive models of science*, pages 3–44. University of Minnesota Press, Minneapolis.

Newell, A. and Simon, H. A. (1976). Computer science as empirical inquiry: symbols and search. *Communications of the ACM*, 19(3):113126.

Niiniluoto, I. (1999). Defending abduction. *Philosophy of Science*, 66:pp. S436–S451.

Niyogi, S. and Snedeker, J., editors (2005). *Symposium on solutions to Fodor's problem of concept acquisition. Twenty-seventh Annual Conference of the Cognitive Science Society.*

Onishi, K. H. and Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308(5719):255 –258.

Piaget, J. (1952). *The origins of intelligence in the child.* International Universities Press, New York.

Piaget, J. (1954). *The construction of reality in the child.* Basic books, New York.

Prinz, J. J. (2002). *Furnishing the mind: Concepts and their perceptual basis.* MIT Press, Cambridge, MA.

Prinz, J. J. (2011). Has mentalese earned its keep? On Jerry Fodor's LOT 2. *Mind*, 120(478):485–501.

Prinz, J. J. and Barsalou, L. (2000). Steering a course for embodied representation. In Markman, A. and Dietrich, E., editors, *Cognitive Dynamics: Conceptual change in humans and machines.* MIT Press, Cambridge, MA.

Putnam, H. (1975). *Mind, language and reality. Philosophical papers vol. 2.* Cambridge University Press, Cambridge.

Putnam, H. (1999). *The threefold cord: Mind, body and world.* Columbia University Press, New York.

Quine, W. V. O. (1953). *From a logical point of view: 9 logico-philosophical essays.* Harvard University Press, Cambridge, MA.

Quine, W. V. O. (1960). *Word and object.* MIT Press, Cambridge, MA.

Quinn, P. C., Eimas, P. D., and Rosenkrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception*, 22(4):463–475.

Quinn, P. C., Schyns, P. G., and Goldstone, R. L. (2006). The interplay between perceptual organization and categorization in the representation of complex visual patterns by young infants. *Journal of Experimental Child Psychology*, 95(2):117–127.

Rosch, E. and Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605.

Rumelhart, D. E. and Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science*, 9(1):75–112.

Russell, B. (1912/1959). *The problems of philosophy.* Oxford University Press, Oxford.

Saby, J., Marshall, P., and Meltzoff, A. (2012). Neural correlates of being imitated: An EEG study in preverbal infants. *Social Neuroscience*, 7:650–661.

Schyns, P. G., Goldstone, R. L., and Thibaut, J.-P. (1998). The development of features in object concepts. *Behavioral and brain sciences*, 21:1–17.

Schyns, P. G. and Murphy, G. L. (1994). The ontogeny of part representation in object concepts. In Medin, D. L., editor, *The Psychology of Learning and Motivation*, pages 305–349. Academic Press.

Schyns, P. G. and Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(3):681–696.

Sellars, W. (1956). Empiricism and the philosophy of mind. In Feigl, H. and Scriven, M., editors, *Minnesota Studies in The Philosophy of Science, Vol. I: The Foundations of Science and the Concepts of Psychology and Psychoanalysis*, pages 253–269. University of Minnesota Press, Minneapolis.

Sellars, W. (1963). Abstract entities. *Review of Metaphysics*, 16:627–671.

Sellars, W. (1968). *Science and metaphysics. Variations on Kantian themes.* Routledge and Kegan Paul, London.

Sellars, W. (1969). Language as thought and as communication. *Philosophy and Phenomenological Research*, 29(4):506–527.

Sellars, W. (1974). Meaning as functional classification. *Synthese*, 27(3):417–437.

Sellars, W. (1980). Behaviorism, language and meaning. *Pacific Philosophical Quarterly*, 61:3–30.

Sellars, W. (1981a). Foundations for a metaphysics of pure process. The Carus lectures. *The Monist*, 64(1):3–90.

Sellars, W. (1981b). Mental events. *Philosophical Studies*, 39:325–345.

Shea, N. (2011). New concepts can be learned. *Biology and philosophy*, 26:129–139.

Smith, C. L. (2007). Bootstrapping processes in the development of students' commonsense matter theories: Using analogical mappings, thought experiments, and learning to measure to promote conceptual restructuring. *Cognition and Instruction*, 25:337–398.

Smith, L. B. and Kemler, D. G. (1978). Levels of experienced dimensionality in children and adults. *Cognitive Psychology*, 10(4):502–532.

Stanford, K. (2009). Underdetermination of scientific theory. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Winter 2009 edition.

Stich, S. and Murphy, D. (1999). Review of Paul Griffiths's What emotions really are. *AAHPSSSS*, pages 13–25.

Stöckle-Schobel, R. (2012). Perceptual learning and feature-based approaches to concepts – a critical discussion. *Frontiers in Psychology*, 3(93).

Strevens, M. (2000). The essentialist aspect of naive theories. *Cognition*, 74(2):149–175.

Thagard, P. (1978). The best explanation: Criteria for theory choice. *Journal of Philosophy*, 75:76–92.

Tomasello, M. (1999). *The cultural origins of human cognition.* Harvard University Press.

Tomasello, M. (2003). *Constructing a language. A usage-based theory of language acquisition.* Harvard University Press, Cambridge, MA.

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(05):675–691.

Tomasello, M. and Farrar, M. J. (1986). Joint attention and early language. *Child Development*, 57(6):1454–1463.

Van Fraassen, B. C. (1977). The pragmatics of explanation. *American Philosophical Quarterly*, 14:143–150.

Van Fraassen, B. C. (1980). *The scientific image.* Oxford University Press, Oxford.

Vygotsky, L. S. (1962). *Thought and language.* MIT Press, Cambridge, MA.

Walden, T. A. and Ogan, T. A. (1988). The development of social referencing. *Child Development*, 59(5):1230–1240.

Wittgenstein, L. (1953). *Philosophical investigations.* Blackwell, Oxford.