Filosofisia tutkimuksia Helsingin yliopistosta
Filosofiska Studier från Helsingfors universitet
Philosophical Studies from the University of Helsinki

Publishers:
Theoretical Philosophy
Social and Moral Philosophy

P.O. Box 24 (Unioninkatu 40A)
00014 University of Helsinki
Finland

Samuli Pöyhönen

# Chasing Phenomena

## Studies on classification and conceptual change in the social and behavioral sciences

Academic dissertation

To be presented, with the permission of the Faculty of Social Sciences
of the University of Helsinki, for public examination in
Auditorium XII, University Main Building, on 5 August 2013,
at 12 noon.

ISBN 978-952-10-9026-4 (paperback)
ISBN 978-952-10-9027-1 (PDF)
ISSN 1458-8331 (series)
Turku 2013. Juvenes Print

# Abstract

The articles comprising this dissertation concern classification and concept formation in the social and behavioral sciences. In particular, the emphasis in the study is on the philosophical analysis of interdisciplinary settings created by the recent intellectual developments on the interfaces between the social sciences, psychology, and neuroscience. The need for a systematic examination of the problems of conceptual coordination and integration across disciplinary boundaries is illustrated by focusing on phenomena whose satisfactory explanation requires drawing together the theoretical resources from a variety of disciplines.

In philosophy, questions regarding the nature of scientific concepts have often been framed in terms of theories of natural kinds. For this reason, analysis of the notion of natural kind as well as examination of how theories of natural kinds should be connected to recent philosophical accounts of scientific explanation and mechanisms form the core of the study. Building on contemporary discussions on these topics in the philosophy of biology, the philosophy of cognitive science, and the philosophy of the social sciences, the articles develop a mechanistic theory of natural kinds in the social and behavioral sciences, and scrutinize its applicability and usefulness as a theory of conceptual change in interdisciplinary settings. The study suggests that, although the mechanistic theory cannot account for the functioning of the whole range of scientific concepts, interweaving biological, cognitive, and social mechanisms – in the manner suggested by the mechanistic theory – offers a naturalistic and non-reductionist basis for conceptualizing epistemic coordination across disciplinary boundaries.

# Contents

# Preface

My intellectual trajectory in the world of academic education has been a convoluted one with several U-turns and loose ends. However, with the benefit of hindsight, there is a certain logic to this path, starting from engineering studies, interrupted by a growing interest in their apparent antithesis, philosophy, and eventually leading to a grey area between the humanities and the sciences, the philosophy of the social and behavioral sciences. As a like-minded colleague once pointed out to me, my philosophical temperament still reflects this background – in both good and bad, my work could often be characterized as philosophy suited for the engineer's mind.

During my work on this doctoral dissertation, I received invaluable help and advice from a wonderful group of people. My greatest thanks go to my supervisor Professor Petri Ylikoski, who invited me to join the Helsinki Philosophy of Science Group, first as a research assistant, and later as his doctoral student. In addition to being a thoughtful and supportive supervisor, Petri's intellectual curiosity, courage, and his ability to see into the heart of the issue at hand have made him an important academic role model for me. Of equal importance, Academy Professor Uskali Mäki's emphasis on conceptual clarity, elegance, and his unerring sense of the dialectic of argumentation have been a source of inspiration and insight during these past seven years that I have been associated with his research group.

Being involved in the projects led by Uskali and Petri, I have had the opportunity to be surrounded by a large group of academic mentors and colleagues. My most sincere thanks go to everyone in the Helsinki Philosophy of Science Group and the Finnish Centre of Excellence for the Philosophy of Social Sciences (TINT) for inspiring discussions, ideas, and the excellent feedback that I have received for my unfinished work. More people than can be mentioned here have helped me by presenting valuable comments on my articles, but I would especially like to thank Marion Godman, Caterina Marchionni, and Jaakko Kuorikoski, who gave excellent comments on an earlier

version of the introductory essay of this thesis. Thanks also to Juho Pääkkönen for word-processing assistance with the essay.

In addition to the co-authored articles included in this dissertation, I have had the honor to write papers together with Tomi Kokkonen and Anna-Mari Rusanen. Through these projects, I have learned that shared intellectual efforts are often far more enjoyable than solitary ones, and tend to result in better outcomes. Furthermore, sharing an office with Aki Lehtinen, Cate, and Jaakko has transformed how I think about the daily life of a researcher: while uninterrupted contemplation no doubt has a role to play in philosophical research, so do the new ideas stumbled upon during lunchtime discussions, and the spontaneous exchange of thoughts over space dividers in the office. My special thanks for intellectual companionship and advice go to Jaakko. Working on shared projects with him has taught me important lessons in analytical clarity, and in how academic papers are written. Without Jaakko's brilliant ideas and our brainstorming sessions, my dissertation project would have been a lot longer and a much less happy one.

In addition to the members of our group, I wish to thank Otto Lappi at the cognitive science unit at the University of Helsinki, and Professor Riitta Hari and the aivoAalto group at Aalto University for offering me a view to the cutting-edge research conducted in the mind sciences. Also the research visits to both the University of Edinburgh and the Berlin School of Mind and Brain widened my philosophical outlook, and during these visits, my hosts Professor Andy Clark and Professor Michael Pauen both kindly offered assistance and comments on my unfinished ideas. I thank Professor Lisa Bortolotti and Professor Gualtiero Piccinini for the pre-examination of this dissertation, and Lisa Muszynski for careful language revision of most of the articles comprising its content (it goes without saying that the remaining errors are my own responsibility). I would also like to thank Professor Matti Sintonen for offering good advice and much needed encouragement in the early stages of my academic career, and Docent Panu Raatikainen who has included me in the research team of several planned projects. Pekka Mäkelä has done a wonderful job in keeping our research community functional, and Karolina Kokko-Uusitalo and

Ilpo Halonen at the office have always offered their assistance when needed. Financially my doctoral research was made possible by a grant from the Finnish Cultural Foundation and a four-year research funding granted by the Finnish Doctoral Programme in Philosophy.

In addition to my academic colleagues, I am also deeply grateful to my other friends and my family. Our musical pursuits together with Artturi Taira, Sampsa Väätäinen, and Jussi Hietala have always been fun, but also a lesson in the confidence and perseverance it takes to make something – be it a song or a scientific publication – so that it is not just 'nearly there,' but exactly like you think it should be. Of course, my thanks go also to my parents Pentti and Seija for unconditionally supporting me in what I have decided to do in life. Finally, the importance of sharing these past years with Johanna and our two daughters goes beyond both my words and my grasp. The time spent in the fun, buzzing world of Inari and Unna has served as a perfect counterpoint to the abstract pleasures and pains of academic work. Finding my way in this whirlwind of a life would not have been possible without Johanna, her dauntless spirit, encouragement, and compassion.

Berlin, 1 July 2013
Samuli Pöyhönen

# List of original publications

This thesis is based on the following publications:

**I** **Carving the mind by its joints: Culture-bound psychiatric disorders as natural kinds**
Samuli Pöyhönen
In K. Talmont-Kaminsky, & M. Milkowski, (eds.) (2013), *Regarding the Mind, Naturally*: *Naturalist Approaches to the Science of the Mental* (pp. 30–48). Cambridge Scholars Publishing.

**II** **Looping kinds and social mechanisms**
Jaakko Kuorikoski and Samuli Pöyhönen
*Sociological Theory*, 30:3, (2012), 187–205.

**III** **Intentional concepts in cognitive neuroscience**
Samuli Pöyhönen
*Philosophical Explorations*, 16:3, forthcoming 2013.

**IV** **Explanatory power of extended cognition**
Samuli Pöyhönen
*Philosophical Psychology*, forthcoming 2013.

**V** **Understanding non-modular functionality – Lessons from genetic algorithms**
Jaakko Kuorikoski and Samuli Pöyhönen
*Philosophy of Science, PSA 2012 proceedings,* forthcoming 2013.

**VI** **Natural kinds and concept eliminativism**
Samuli Pöyhönen
In K. Karakostas, & D. Dieks, (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, Dordrecht, Springer, forthcoming 2013.

The publications are referred to in the text by their roman numerals.

# Part I: Introductory essay

## 1. Introduction. Phenomena on interdisciplinary boundaries

The six papers comprising this dissertation concern classification and concept formation in the social and behavioral sciences, i.e. in disciplines whose principal aim is to study human behavior and its causes.[1] In particular, I aim to shed light on a scientifically important but philosophically rather underdeveloped topic: conceptual change in interdisciplinary settings. I approach this topic by employing philosophical theories of concepts, natural kinds, and scientific explanation to develop a *mechanistic theory of natural kinds* in the social and behavioral sciences.

To get started, consider the multiple disciplinary perspectives involved in the practices surrounding a psychiatric phenomenon, *Asperger's syndrome*. On the one hand, diagnostic criteria of the disorder as well as important treatments of autism spectrum disorders (e.g., psychosocial interventions) rely on common-sense psychological descriptions of the phenomenon (Roth & Barson 2010, ch. 2, 6). On the other hand, theoretically the most unified approach to autism spectrum disorders is to describe them as deficiencies in the cognitive and attentional capacities constituting the theory of mind (Baron-Cohen 2000; Premack & Woodruff 1978). Moreover, in addition to scientific psychology, also neuroscientific, genetic, and computational approaches have recently provided interesting insights into the explanation of psychiatric disorders like Asperger's (Arguello & Gogos 2012; Brock 2012; Montague et al. 2012).

---

[1] Among the social and behavioral sciences, I include the social sciences, the different branches of psychology including biological psychology, and linguistics as well as cognitive science (Smelser & Baltes 2001). The boundaries of this group of disciplines are not sharp, and for the purposes of this dissertation, I also include psychiatry and large parts of medicine within this disciplinary constellation.

Like many other phenomena studied in the social and behavioral sciences, Asperger's is not only a target of scientific research but also a societally important and value-laden classification. For this reason, research findings on the disorder have provoked reactions in those classified, and the changing behavior has thus made the phenomenon somewhat unstable (Eyal 2010). The concept has also, in a sense, leaked into the surrounding population and created practices of lay-diagnosis characterizing also shy and perhaps socially awkward people in terms of autism spectrum disorders. Consequently, regardless of the correctness of these diagnoses, the classification and the scientific knowledge pertaining to the phenomenon have, in an interesting sense, participated in constructing social reality (Hacking 2002, ch. 6).

What the existence of the multiple conceptualizations of Asperger's syndrome suggests is that it, in a way, resides on an intellectual boundary between various different scientific fields, and it also exemplifies the conceptual traffic between science, everyday thinking, and policy. This multi-perspectival nature is characteristic also of many of the other examples examined in this dissertation. The research on human psychological abilities such as concepts, memory, or vision, and the study of culturally embedded psychiatric phenomena, like eating disorders, all raise different challenges to scientific theorizing, but what draws these examples together is the need for *interfield integration*: comprehensive understanding of these phenomena requires bringing together theoretical contributions from a variety of disciplinary perspectives.

Different fields often study roughly the same target phenomena at various levels of abstraction by using different theoretical and material tools, and as a result, end up with apparently very different conceptualizations of the same target phenomena. Hence, a central problem that must be answered in order to achieve conceptual coordination across disciplinary boundaries is *how to connect the different ways of conceptualizing the phenomenon*. In 20th-century philosophy, questions of interdisciplinary organization were often framed in terms of two compet-

ing alternatives. On the one hand, reductive models of the unity of science (Nagel 1961; Oppenheim & Putnam 1958) suggest that higher-level concepts and theories can ultimately be derived from lower-level ones. On the other hand, those advocating the autonomy of the special sciences have alluded to theoretical, ontological, or methodological discontinuities between disciplines as reasons for insulating research in higher-level disciplines from lower-ones, thus safeguarding them from reduction (Fodor 1974; Dupré 1995; Cartwright 1999; Geertz 1973; Taylor 1971).

In the light of recent developments in the social and behavioral sciences, both these alternatives seem inadequate. On the one hand, epistemic collaboration across disciplinary boundaries is a fact of the matter in science. For example, results regarding the psychology of decision-making and social cognition have increasingly been used to provide a more realistic psychological basis for theories in the social sciences (Elster 2007; Evans & Frankish 2009; Gigerenzer, Hertwig, & Pachur 2011; Kahneman 2011; Sperber 1996). Also recent advancements in neuroscience and genetics have shown that these natural-science disciplines are exciting sources of evidence for the psychological sciences, and in experimental psychology, it has become commonplace to support findings by referring to evidence achieved by neuroscience methods. In the face of these developments, claims of strong disciplinary autonomy or theoretical disconnectedness of disciplines from each other appear implausible.

However, unlike the reductionist unity view, these interdisciplinary encounters do not usually suggest the reduction of higher-level theories and concepts to lower-level ones. Even the contemporary defenders of reductionist models agree that genuine reductions in science are rare (Craver 2007, 233). Instead, multi-disciplinary research programs in the sciences often conceive of different fields as complementary sources of evidence for theorizing (Cacioppo & Decety 2011; Glimcher & Rustichini 2004). Moreover, conceptual influence between levels is not always directed from lower-level theories to the higher-

level ones. Often understanding micro-level structures in the mind sciences requires making clear the larger-context in which they function (Bechtel 2009). For example, social and psychological sciences can influence lower-level ones by suggesting new explananda, and by making traditional explananda more precise by describing the social and ecological environmental challenges that our psychological capacities might be evolved to meet (Duchaine, Cosmides, & Tooby 2001; Dunbar 1998; Laland & Brown 2002; Sterelny 2003).

These interdisciplinary developments both within the social and behavioral sciences and across the boundaries of this group of disciplines reflect the fact that evidential and explanatory relationships do not respect disciplinary boundaries, and the differences in the ways that scientific disciplines conceptualize target phenomena do not make causal interrelations go away. Therefore, the challenge faced by philosophical accounts of interdisciplinary organization is to show how it would be possible to move beyond mere disciplinary pluralism towards a view that can satisfactorily account also for interfield coordination and integration.

The papers comprising this dissertation address the problems of concept formation and conceptual change in interdisciplinary settings by building on recent discussions on natural kinds and scientific explanation in the philosophy of science. I suggest that although phenomena studied in the social and behavioral sciences often transgress disciplinary boundaries in the sense that there is no one correct disciplinary perspective for their satisfactory description and explanation, these phenomena can still often be conceived of as natural kinds – or at least as scientific kinds. In particular, I propose that a mechanistic theory of kinds building on Richard Boyd's homeostatic property cluster theory offers a workable platform for conceptualizing phenomena on interdisciplinary interfaces. My three general contributions to the existing literature can be summarized as follows:

(1) The articles examine how the mechanistic theory of kinds, originating in the philosophy of biology, could be applied in the social and behavioral sciences, where phenomena often involve intentional action, social processes, and complex causal interactions between personal and subpersonal factors. I argue that elaborating the mechanistic theory, by connecting it to recent accounts of scientific explanation and mechanisms, results in a naturalistic but non-reductionist picture of concepts and phenomena that can meet this challenge. My approach suggests that strict divisions of phenomena into "social" and "natural" are often artificial and a hindrance to interdisciplinary knowledge production, and I argue that the mechanistic theory of kinds can shed light on classificatory controversies surrounding topics such as social construction and extended cognition.

(2) As a dialectical counterpart to the first contribution, my analysis also points to the limitations of the mechanistic approach. First, I argue that in the causally complex domains studied by the social and behavioral sciences, the mechanistic theory of natural kinds often does not provide unanimous taxonomies of fields of phenomena. Instead, it recommends classificatory pluralism, where the choice of the units of analysis in a scientific discipline is determined both by the causal structure of the world and by the epistemic aims of the discipline. Consequently, it is possible for causally-based classifications in different scientific fields both to reside at widely different levels of abstraction and to cross-classify.

Furthermore, as a more fundamental challenge to the mechanistic approach, the articles also hint towards insufficiency of what could be called *the natural kinds model* of scientific concepts. Paying attention to concept use in scientific research practices suggests that not all concepts can be aligned with causal structures, but instead several indispensable scientific concepts are more general terms that function as devices for interdisciplinary coordination. Hence, it is reasonable to be suspicious of the idea of all scientific concepts functioning in the same way.

(3) The third main contribution goes some way towards resolving the tension between the two previous results. I suggest a slight reinterpretation of the common view of natural kinds in the sciences. Instead of approaching natural kinds as the fundamental metaphysical building blocks of reality, I view kind terms used in science as *epistemic tools* that the community of cognitively finite agents employs in studying causally complex phenomena. Furthermore, I propose that such epistemic tools come in different varieties, and the semantic dissimilarities between different types of scientific concepts reflect their different epistemic roles in research. Finally, building on these results obtained in the articles, in the last section of this introductory essay, I outline a mechanism-based picture of conceptual integration in interfield settings, in which accumulation of knowledge across disciplinary boundaries is achieved by uncovering *inferential constraints* that parallel theories in neighboring fields impose on each other.

The structure of this introductory essay is as follows. In section 2, I provide a tentative roadmap of theories of concepts and natural kinds that can be used to contextualize the (perhaps slightly unorthodox) notion of scientific concept employed in the articles. This stage-setting is done in two steps. First, I argue that theories of concepts in psychology and philosophy should often be seen as two different theoretical projects. Second, by reviewing ways of theorizing about natural kinds, I show that there is an important and somewhat self-sustained strand of the philosophical discussions on natural kinds that has been primarily motivated by concerns of the epistemic reliability of scientific concepts. The aim of the resulting roadmap of theories of concepts and natural kinds is to suggest that the account of scientific conceptual change developed in articles is not directly constrained by many philosophical discussions on these notions outside the philosophy of science (e.g., in metaphysics or the philosophy of mind).

Building on this groundwork, in section 3 I show how Richard Boyd's theory of natural kinds as homeostatic property clusters pro-

vides a promising starting point for developing a view of the nature and the functioning of scientific concepts in the social and behavioral sciences. In section 4, I introduce the theoretical resources from recent discussions on scientific explanation and mechanisms that are utilized in the articles. I argue that the contrastive-counterfactual theory of explanation (Woodward 2003; Ylikoski 2001) together with recent discussions on mechanistic research heuristics (Bechtel & Richardson 2010) offer crucial conceptual tools for turning the mechanistic theory of kinds into a more comprehensive picture of concept formation in interdisciplinary settings.

After presenting the theoretical background in sections 2 to 4, section 5 provides an overview of the articles, and I conclude the introductory essay by sketching a general outlook on conceptual change in the social and behavioral sciences suggested by the results obtained in the articles.

Finally, before setting off, a word on the method employed in the studies. Most of the articles in this dissertation discuss case studies drawn from various social and behavioral sciences: multi-disciplinary cases from the cognitive sciences play a central role, as well as ones from economics, sociology, and psychiatry. The case-based approach reflects my naturalistic approach to the philosophy of science – the conviction that the philosophy of science is best done in close contact with actual research in the sciences. Although this requires dividing one's time between philosophical literature and that of the target sciences, it also opens up new interesting philosophical questions, and hopefully sometimes produces results useful also for the scientists.

Despite the engagement with empirical research, the articles are philosophical in nature, and the main method employed is conceptual analysis, understood in a very broad sense. In other words, the papers do not introduce new empirical results, but instead my aim is to understand how scientific concepts are used in various contexts, and how these usages are related to each other (Griffiths & Stotz 2008; Hacking 2002, 24–25). Moreover, my perspective on conceptual

change is not that of a historian or sociologist of science, but instead I use the cases for trying to illustrate general properties of the conceptual dynamics in scientific research.

In analyzing scientific materials, I understand the role of a philosopher as that of an inferential scorekeeper: by making it clear how concepts are used in different research contexts, a philosopher can sometimes act as mediator, whose goal is to try to see the bigger picture by understanding how things generally hang together (Sellars 1963). When it comes to questions of interfield cooperation, such contributions appear particularly useful, because, based on my own impression, not all scientists have a well-calibrated picture of the relevance of their work outside their own discipline, nor do they often have a systematic view of how to think of epistemic coordination or collaboration across disciplinary boundaries. Furthermore, a clear picture of the relevance of scientific knowledge across disciplinary boundaries would be important also when scientific results are applied in solving practical societal problems.

## 2. Scientific concepts

Study of scientific concepts and conceptual change forms the core of this dissertation, and therefore I should begin with some terminological remarks concerning my use of these notions. In the philosophical and psychological literature, 'concept' is typically taken to refer to a mental, linguistic, or perhaps even abstract entity that represents a 'kind' or 'category', something in reality corresponding to it (cf. Murphy 2004, 5). Although disagreements reign over the nature of both sides of this representational relation, the distinction between the representation and its target is clear enough, and I follow the entrenched usage of these notions.

Respectively, by 'conceptual change,' I refer to theoretical developments in the sciences, as they are reflected in the use and contents of scientific concepts. I understand conceptual change in a broad sense that captures both concept formation and conceptual revision in

the sciences, and by using the phrase, I mainly intend to convey that my study of epistemic change proceeds at the level of concepts rather than, for example, on the level of theories, models, or paradigms. Unlike it is often assumed in conceptual change research in fields such as science education or developmental psychology, I do not presuppose a strong analogy between conceptual change in individual learning and scientific theorizing. Moreover, my use of the phrase should not be taken to imply that conceptual change should always involve discontinuities or radical reorganization of conceptual schemes (Kuhn 1996; Vosniadou 2008). By contrast, I suggest that the study of conceptual change in science should involve keeping careful track of the often subtle changes in the inferential potential of scientific concepts (cf. Griffiths & Stotz 2008).

## 2.1. Focusing on concepts

Although examining the level of concepts and conceptual change is by no means the only possible approach to studying theoretical dynamics in the social and behavioral sciences, I hold that there are several reasons why it is an interesting perspective on the epistemic dynamics of interdisciplinary research.

First, in philosophy, psychology, and in the methodological discussions in the sciences, concepts are regarded as playing a fundamental role in inductive inference. Psychologically, concepts function as the mental glue that connects our prior knowledge of kinds of things in the world to new targets that we encounter. They enable us to see repetition, sameness, and continuity in the diverse world, and therefore help us fit the infinite reality to our limited cognitive capacities (Smith & Medin 1981, 8). Likewise, in philosophy, discussions on inductive inference have often been framed in terms of classes, kinds, or categories and their properties that enter into inductive generalizations (Aristotle, *Topics*; Goodman [1955]1983; Mill [1891]2002). The methodological discussions in the sciences on operationalization and construct validity (Cronbach & Meehl 1955; Shadish, Cook, & Camp-

bell 2002, ch. 3) also reflect the central inferential role of concepts. Regimenting the correct use of theoretical concepts is essential for ensuring reliable knowledge production in research.

From all these perspectives, concepts license inferences by imposing structure on perception or content. This fact makes the question of how to formulate our concepts an important normative issue: generalizations we make for prediction, explanation, and manipulation of phenomena rely on our conceptualizations of such phenomena.

In addition to this role in grounding inductive inference, both philosophers and sociologists have pointed out the *coordinative role* that central scientific concepts have in research. Multi-faceted notions such as GENE, ENERGY, and RATIONALITY feature simultaneously in multiple theories and have functioned as nodes around which successful research has been conducted (Bermudez 2004, ch. 1; Putnam 1975a; Star & Griesemer 1989).

In philosophy, the *tradition of natural kinds* since John Stuart Mill (2002) has focused on these scientifically central concepts, and the literature on kinds suggests an explanation for the focal role of scientific concepts and classifications. Their epistemological centrality appears to follow from the fact that the choice of the units of analysis within a field of research often reflects scientists' judgments of *where the genuine phenomena are*. It is a widely shared ideal that our classificatory concepts should divide the domain of study in a way that reflects the objective structure of reality, and thus makes reliable epistemic practices possible.[2]

---

[2] By focusing on classificatory concepts, I do not mean to imply that all scientific concepts are primarily classificatory (think of DEMOCRACY, CULTURE, or MASS) (cf. Millikan 2000, §1.1). However, in this study I examine concepts whose main role in thought and theory is to pick out individual targets in the world, and to lump them together with others to gain epistemic benefits. Moreover, I use the notion of classification in a broad sense, and I do not require that scientific classifications must be exclusive, exhaustive, hierarchical, or non-cross-cutting (cf. Hacking 1993). To the

Finally, the debates on where the genuine phenomena are also have important pragmatic and policy implications. Medical, psychiatric, and social scientific classification-systems such as ICD and DSM play an important role in policy-making and are often targets of intense debates. Whether Asperger's syndrome, eating disorders, addictions, or post-traumatic stress disorder are regarded as genuine illnesses has crucial implications for practical issues such as treatment interventions and insurance coverage. Therefore, having a philosophical theory of the nature of such concepts and classifications would be highly desirable.

Thanks to the central role of concepts both in psychology and philosophy, the literature on the topic in both disciplines is extensive. I will not engage in the daunting task of trying to review these discussions.[3] However, in the rest of this section I provide a tentative roadmap to theories of concepts and natural kinds, which can be used to contextualize the notion of scientific concept employed in the following articles. My approach to scientific concepts is based on two distinctions. First, I suggest that psychological and philosophical theories of concepts often arise from distinct theoretical concerns, and the psychological and semantic perspectives on concepts should be clearly distinguished from each other.[4] Secondly, I argue that in the philosophy of science, the question concerning the nature of scientific concepts is the primary motivation for theories of natural kinds. I hold that the discussions on natural kinds in other philosophical fields such as metaphysics and the philosophy of language are often tangential to

contrary, as is suggested in sections 5.5. and 6.2., classificatory systems in the social and behavioral sciences rarely satisfy such stringent criteria.

[3] For good book-length treatments of concepts in philosophy and psychology, see e.g. Murphy 2004; Margolis & Laurence 1999.

[4] A similar suggestion has been put forward by Edouard Machery (2009, ch. 2). In what follows, I partly rely on Machery's insightful treatment of the topic.

the concerns of the *epistemology-oriented tradition of natural kinds* in the philosophy of science.

## 2.2. Concepts in psychology and philosophy

In psychology and the cognitive sciences, conceptual thought is conceived of as a psychological capacity that (at least) we humans have. Concepts are understood as mental representations that go beyond mere sensory content, and are capable of representing absent and sometimes abstract targets. The central theoretical aim of psychological concept theories has been to explain how these cognitive feats are realized by the functioning of cognitive and neural mechanisms. More precisely, the goal of psychological theories of concepts has usually been to determine what kind of knowledge features in processes underlying the higher cognitive capacities, how this knowledge is used in these processes, how it is acquired, and where it is located in the brain. By doing so, the theories hope to explain various properties of cognitive competences such as categorization, inductive inference, memory, learning, problem-solving, analogical inference, and language.

Such explanatory projects typically conceive of concepts as information structures stored in long-term memory (Machery 2009, 12). Psychologists are by no means unanimous about the nature of these structures, and disagree whether concepts are property clusters, exemplars, theory-embedded structures, or perhaps *ad-hoc* compositions created on the fly, as has been suggested by concept empiricists (Barsalou 1999; Prinz 2002). Despite these deep-seated disagreements, psychological theories typically share an ontological view of concepts as real psychological structures that can be referred to in causal explanations of individual behavior.

Among the wide array of modern philosophical theories of concepts, there appear to be only two fundamental properties that the theories mostly agree on: concepts are a way to *refer* to targets in the world, and they carry *content* regarding the properties of those targets. Theories of

conceptuality fail to agree on much else. Firstly, theories of neurosemantics, mental representation, linguistic concepts, and scientific concepts all approach content at *different levels* of inquiry (Eliasmith 2005; Dretske 1981; Fodor 1998; Piccinini & Scott 2006; Churchland 1984, 56). Secondly, atomistic theories and those advocating inferential role semantics strongly disagree on how much inferential content and connections to other concepts should be included in the meaning of concepts (Fodor 1998; Block 1998; Brandom 1994). Thirdly, a sizable industry in the 20th century philosophy of mind and language focused on the disagreement as to whether the reference of concepts is fixed by causal contact with their correspondents in reality, or by the cognitive content of concepts (Frege [1892]2010; Kripke 1980; Putnam 1975b).

Despite these major differences, it is possible to distinguish common characteristics of the philosophical theories of concepts that set them apart from the psychological ones. In stark contrast to the approach in psychology, for the last hundred years, the philosophical tradition has been strongly anti-psychologist, and has often emphasized the sharedness and abstract nature of concepts. From the paradigmatic philosophical viewpoint, concepts are not understood as psychological structures but rather as *constituents of propositional contents* (Laurence & Margolis 1999, 6), and conceptual identity is not determined by referring to causally real information structures in the head, but instead by operations on conceptual content. For instance, in Fregean spirit, Christopher Peacocke (1992, 2) suggests that two concepts are distinct if substituting one for the other in an otherwise identical proposition could create cognitive novelty.

I take it that these properties reflect the fact that philosophical discussions of concepts primarily concern their semantic properties. Typically, the philosophical theories aim to specify the conditions that have to hold in order for a concept to refer to a particular group of things in the world. For example, descriptivism, causal theories of reference, and teleosemantic theories all offer different accounts of

the nature of the relationship between a representation and the set of things in the world that it correctly applies to. As noted by Jussi Jylkkä (2008, 10), theories of concepts are often assumed to "*determine not only which objects we reckon as belonging in categories, but also which objects truly belong in them*." Importantly, this question is quite alien to the psychological theories, and therefore inability to provide an account of the possibility of misrepresentation, or answer issues raised by ignorance and error (Putnam 1975b), should not count as arguments against the adequacy of psychological theories.

In sum, one could say that whereas psychological theories are ultimately in the business of *explaining* the functioning of cognitive structures, semantic theories in philosophy aim to *explicate* the conditions that would need to hold for someone to possess a certain subpropositional thought content.

Of course, the differences between these viewpoints do not mean that there would be no interesting relationships between the psychological and philosophical theories. In classical and early modern theories of concepts, no clear distinction was often made between the two viewpoints (Locke [1690]1975; Mill 2002). Also after Frege, the coevolution of philosophical and psychological theories has continued. The development of psychological theories starting from the classical view of concepts (Hull 1920) to prototype theory (Rosch 1978; Rosch & Mervis 1975;) and theory-based approaches (Gopnik & Meltzoff 1997; Murphy & Medin 1985) were strongly influenced by parallel philosophical developments (Carnap 1932; Putnam 1975b; Wittgenstein 1953). Recently, there has also been conceptual traffic in the opposite direction: often the contemporary philosophical theories of concepts and natural kinds summarize findings on concept structure in psychology, and compatibility with psychological theories is sometimes used as an argument in favor of a particular philosophical theory (Griffiths 1998, 176–192; Kornblith 1993, ch. 4).

In the philosophy of mind and the philosophy of psychology, the relationship between these two points of view has often been seen as particularly intimate, and the theories of concepts in these fields sometimes fall between what I described as the psychological and semantic viewpoints. A well-known example of the approach closely linking them is Jerry Fodor's theory of concepts (1998; 2010). According to Fodor, concepts serve a dual role as real entities in the language of thought and also as constituents of propositional contents; language of thought is the causally real platform for propositional thoughts, thus acting as a bridge between the two perspectives. Importantly, a similar presupposition sometimes underlies empirical research in the cognitive sciences. For example, Susan Carey's psychological theory of conceptual change attempts to remain sensitive to philosophers' requirements for a theory of concepts (Carey 2009, ch. 13).

These hybrid theories reflect the methodological conviction in both the philosophy of mind and the philosophy of psychology that a theory of concepts should be able to answer questions arising from both psychological and semantic perspectives by relying on one unified theory of the topic. For instance, in their influential reviews of theories of concepts, Stephen Laurence and Eric Margolis pit the psychological and philosophical theories against each other, and assess each in the light of the same set of challenges allegedly faced by theories of concepts in general (Laurence & Margolis 1999; Margolis & Laurence 2012).

It seems that ultimately the rationale for the hybrid theories is a general commitment to naturalism: creating a close connection between the structure of our cognitive machinery and propositionality is often thought of as a precondition for a naturalistic account of propositional thought and its properties (i.e. systematicity, productivity). However, given the important differences between the psychological and semantic concept theories, I do not believe that such a straightforward way of making this connection is helpful. Psychological and semantic theories address largely different questions, and draw on very

different kinds of evidence. Therefore, I find problematic the assumption that a single theory of concepts would be adequate for both these theoretical projects (Rusanen & Pöyhönen 2013). Based on my personal experience of grappling with the often confusing literature on concepts, I have come to believe that for the sake of clarity, it is useful to assume that the notion of concept works in somewhat different ways in the psychological and philosophical literatures. Hence, for the purposes of the current study, I suggest treating psychological and philosophical theories as two distinct *perspectives on concepts*.

That said, I do think that examining the relationships between the psychological and semantic theories in detail would be an interesting task, and that the kind of naturalistic theory of scientific concepts developed in this dissertation should pay attention to considerations arising from both these perspectives. However, such a project should not start by simply collapsing the distinction between them. Instead, having a clear view of the theoretical motivation of each cluster of theories helps to make clear the kind of input that these theories could offer to the kind of theory of scientific concepts developed here.

In the following sections on natural kinds, and in the articles themselves, I will have more to say about the semantic properties of scientific concepts. However, discussions of psychological concept theories or hybrid theories do not play a major role in this dissertation. This is because I do not think that existing psychological research on concepts or the theories in the philosophy of mind could offer very substantial constraints to my account of scientific conceptual change. Firstly, the research on concepts in experimental psychology has often focused on concepts that are as divorced from outside knowledge as possible. It is not entirely clear how the study of non-theoretical concepts employed by isolated human individuals in psychological laboratory experiment settings should influence our view of the cognitive structures employed in complicated scientific inference.

Secondly, a reason that further limits the usefulness of the existing psychological concept theories for my account of conceptual change

has to do with the social aspects of scientific research. Scientific inquiry is a collective process carried out by a community of scientists, and the correctness of scientific inference is often guaranteed by communal practices of communication and error correction. Hence, for the study of scientific concepts it seems that investigating the social processes behind concept transmission in scientific communities would be as important an empirical task as the research done in individualist psychology (Bishop 2002; Downes 1999; Faucher et al. 2002).[5]

For similar reasons, I suspect that many classic accounts in the philosophy of mind of how mental symbols succeed in referring to their targets (Dretske 1981; Fodor 1998) might be largely peripheral for the questions studied in this dissertation. Due to the theoretical nature of scientific concepts and the division of linguistic labor within the scientific community (Putnam 1975b) teleosemantic and information-theoretic accounts do not appear as adequate semantic theories for scientific concepts.

## 2.3. Concepts and reality. Theories of natural kinds

The distinction between the psychological and semantic perspectives on conceptuality is the first step in navigating among the variety of theories of concepts. Moreover, I suggested above that theories of concepts both in psychology and in the philosophy of mind often fail to capture important aspects of scientific concepts. In this section, I

[5] The social nature of scientific theorizing could actually be seen as an even more serious argument against making a theory of scientific concepts conform to a model of concepts originating in psychological research. For example, although Kornblith (1993, ch. 4) and Griffiths (1998, 176–192) appeal to our tendency to *psychological essentialism* (Gelman 2003) as an argument for the theory of natural kinds that they defend, why could it not be so that good scientific concept formation requires intentionally countering certain psychological biases in our thinking (e.g., essentialism) – perhaps good scientific concepts are sometimes psychologically unnatural.

try to shed light on their nature by discussing two questions that a theory of conceptual change in science should be able to answer: (1) Where the epistemic reliability and usefulness of scientific concepts come from, and (2) how, if at all, they are related to structures in reality. These interrelated questions address perhaps the most central concerns in the discussions on *natural kinds* in the philosophy of science. I introduce three influential approaches to answering these questions: essentialism, empiricism, and scientific realism.[6]

The discussions on natural kinds in the philosophy of science represent only one part of the usage of the notion, and the topic has recently received a lot of attention also in other parts of philosophy. In the philosophy of language, natural kinds have played a central role in arguments against descriptivism, and in metaphysics, the concept features in discussions concerning the laws of nature, natural necessity, and essentialism (cf. Bird & Tobin 2008). The purpose of the following selective history of the concept of natural kind is to suggest that in the philosophy of science, theories of natural kinds have largely been motivated by a concern different from the ones above: the *inferential reliability* of scientific concepts. In the articles, I take this idea further and argue that the discussions on natural kinds in the philosophy of science forming what can be called the *epistemology-oriented tradition of natural kinds* (Reydon 2009; cf. Brigandt 2010) are at least partly independent from the other uses of the notion in the philosophy of language and in metaphysics.[7] Consequently, while my approach to natural kinds in the social and behavioral sciences might appear counter-intuitive to some, I suggest that many of the intuitions against my view stem from these other discussions, and should not be under-

[6] The following overview of the tradition of natural kinds is partly based on Ian Hacking's (1991; 2006) seminal research on the topic.

[7] In addition to the separate uses of the term in philosophy, in psychology 'natural kind' is often used to refer to classifications that are psychologically primary or perceptually salient (Berlin 1992; Keil 1992). In this thesis, I do not discuss this psychological use of the notion.

stood directly as grounds for arguments against my theory of classification and conceptual change.

*Aristotelian essentialism*

In philosophy, perhaps the most long-lived picture of the relationship between scientific concepts and reality is Plato's image of "carving nature by its joints" (Plato, *Phaedrus,* 265a–266a). However, in the history of science, a botanical metaphor has played a more pervasive role. Based on an Aristotelian picture of the relationship between nature and scientific research, until the turn of the 19th century, *Porphyry's tree* was the analogy for scientific knowledge. According to the Aristotelian ideal, it was seen as the objective of science to create a universal hierarchy of things by correctly categorizing all things in nature according to their *species* and *genus*. It was thought that each species has a group of properties flowing necessarily from its essence, other properties being accidental (Ayers 1981, 248–252). Scientific knowledge was supposed to correspond to the hierarchical structure of reality, and the inferential legitimacy of scientific concepts was anchored in this correspondence.

In early modern philosophy, this picture of the concurrence between the structure of reality and our conceptual system was challenged by materialist metaphysics and the new empiricist epistemology. John Locke (1975) famously argued that the properties of various natural phenomena are not differentiated by species-specific Aristotelian forms, but instead determined by different configurations of the same material substance. Locke gave two separate arguments for the nominalist approach to classification that he propounded. According to the metaphysical argument, phenomena in nature differ from each other in various ways and hence there is no natural way to classify them. Nature has no single set of joints. The second argument was epistemological: Locke proposed that even if there were microstructural real essences in nature, they would be beyond our perceptual capacities. Consequently, our classifications cannot follow reality's

own divisions but are only the work of our understanding (Locke 1975, III, vi, §36–37).

Despite the empiricist critique of Aristotelian essentialism in philosophy, classification in many of the sciences still relied on the Aristotelian model until the turn of the 19th century. At that time, especially biological classification was in its golden age: Carl von Linne's system of biological nomenclature and categorization offered the theoretical basis for classifying unknown species of plants and animals arriving to Europe from the colonies (Hacking 2006). While Linne's method was based on empirical observation, it still relied on the hierarchical Aristotelian picture of reality consisting of nested species and genera. However, the birth of the theory of natural selection with its emphasis on population-level thinking undermined the philosophical basis for the intuitively appealing Aristotelian essentialism. Within population thinking, there is no explanatory role for a species essence. Explanations of the common properties of species are provided by referring to factors such as common ancestry and evolutionary selection pressures (Grene & Depew 2004, 35–53; Sober 1980).

*Empiricism: The birth of 'natural kind'*

Dispensing with Aristotelian metaphysics made Locke's view an appealing contender for an empiricist approach to the foundations of classification. However, the problem with his view was that it could not give a satisfactory explanation to why some classifications appear more significant than others. Some concepts appear to pick out preexisting divisions in reality, whereas others are based on mere convention and linguistic agreement. Classifications of elementary particles seem more fundamental than classification of people according to their political views, and classifying animals according to their species more natural than grouping them according to the color of their fur. The birth of the notion of *natural kind* was motivated by this tension in empiricism between a nominalist approach to classification and the need to justify inductive inference (Boyd 1991). The notion has been

used to draw the distinction between classifications that merit scientific inquiry and merely conventional classifications, without relying on the metaphysically suspect notion of essence.

Formulating a theory of naturalness of classifications compatible with the nominalist spirit of empiricism was the task taken up by philosophers such as William Whewell ([1847]1967) and John Stuart Mill (2002). As Mill's theory is considered to be the first actual theory of natural kinds, and because it nicely captures the epistemic motivation for the use of the notion, I'll briefly describe it here.

According to Mill, the common names in our language can be divided into two groups. There are (i) words whose referents have only one property in common, and (ii) those whose referents share an innumerable amount of properties that are logically independent of each other. Mill called these groups *finite classes* and *real kinds*, respectively (Mill 2002, I, vii, §3–8). Mill's example of a finite class was the group of white things. The members of this group have nothing else in common than the property of whiteness and others that logically follow from it (e.g., being colored, not-being-red). His examples of real kinds, on the contrary, are classifications that have thereafter often been treated as paradigmatic examples of natural kinds: classes of animals, plants, minerals, and chemical compounds such as sulfur and phosphorus. What the set of real kinds has in common is that scientific research has repeatedly revealed new surprising similarities between the instances of these kinds.

Mill offers the empiricist tradition the first picture of how scientific concepts are special. Without making any deep metaphysical commitments to the existence of kinds in nature, Mill's distinction tries to make precise the convincing intuition that some of our concepts capture true regularities in reality, whereas others may be based on linguistic conventions only. Real – or natural – kinds are the ones that can support reliable inductive inferences and therefore deserve scientific attention.

After Mill, the notion of natural kind drifted away from the limelight for almost half a century. The development of formal logic as a new powerful tool of conceptual analysis, and the verificationist semantics of logical empiricism inspired attempts to reduce the notion of natural kind to less problematic or empirically better-defined notions. For example, Carnap (1950) tried to define kinds in terms of classes – later to be famously criticized by Goodman (1983). Often in these discussions, the notion of kind was connected to other disreputable notions such as disposition and similarity, and both Russell (1948, 462) and Quine (1969) suggested that in mature branches of science, these psychologically salient but scientifically vague notions would be replaced by descriptions of laws or mechanisms underlying the kind dispositions.

In practice, the empiricist discussions on inductive inference often focused on the lawlikeness of hypotheses, and natural kinds were simply thought of as predicates that enter into the law-statements (Fodor 1974; Goodman 1983; Putnam 1975a).[8] However, a shift towards more substantive notions of kind was initiated when it turned out that determining the genuine lawlikeness of hypotheses might require constraining the group of predicates that can enter into such generalizations. Nelson Goodman's grue paradox played a central role in this transition (Stalker 1994). A widely accepted implication of the paradox is that syntactic properties of competing generalizations (involving 'blue' vs. 'grue') alone are not sufficient for distinguishing scientifically legitimate hypothesis from problematic ones, and Goodman himself suggested that gruesome hypotheses could be avoided by setting non-syntactic criteria for the predicates that can feature in these generalizations (Goodman 1983, ch. IV). According to Goodman, genuinely lawlike hypotheses must be phrased in terms of

[8] The tight connection between laws of nature and natural kinds is still endorsed by *nomological theories of natural kinds* (Rosenberg 2005; Hull 1980; cf. Murphy 2006, 336).

*projectible* predicates, where projectible predicates must refer to relevant or genuine kinds. While Goodman's own construal of kindhood in terms of the notion of entrenchment remained firmly within the empiricist approach to kind-systems, a popular approach to articulating what the relevance or naturalness of kind terms could be has been to suggest that natural kind concepts refer to causally real targets in reality. In the following, I call this group of theories *natural kinds realism*.

*Scientific realism and natural kinds*

The emergence of natural kinds realism in the second half of the 20th century was closely related to problems of empiricism, and the consequent rise of scientific realism. The criticisms against a strong analytic-synthetic distinction had questioned the semantically foundationalist project, in which the empiricists hoped to reduce theoretical kind-terms back to observational or physical base vocabularies. In contrast, the realist commitments to the existence of theoretical entities and to the possibility of their accurate representation are reflected in treating theoretical terms as referring expressions anchored in the (often non-observable) causal structures of reality.

Moreover, the emergence of realist interpretations of natural kinds in the philosophy of science was facilitated by Kripke's and Putnam's work in the philosophy of language in the 1970s. Firstly, causal theories of reference suggested a picture of how our natural kind concepts could refer to the same targets in reality despite the sometimes pervasive change in our beliefs regarding them. The possibility of such referential stability has been an influential response to the theory-ladenness and incommensurability arguments underlying the constructivist anti-realist alternatives to scientific realism. Secondly, while their work concerned primarily the philosophy of language, Kripke and Putnam often drew their examples of natural kinds from science, and inspired a metaphysical picture of natural kinds that has been called *simple essentialism* (Murphy 2006, 335–338). Borrowing ways of speaking from the Aristotelian and early empiricist views of science, simple

essentialism suggests that there are natural kinds like GOLD or WATER whose "real essences" are constituted by their microstructure. While the instances of such kinds are often identified by more superficial stereotypical properties, the (sometimes unknown) true extension of the kind is determined by nature alone.

An important advantage of natural kinds realism is that, unlike their empiricist alternatives, realist theories offer a strategy for explaining the property correlations postulated by natural kind concepts by referring to the causal structures underlying them. By the same token, these realist theories can provide a compelling justification for the inductive reliability of natural kind terms (Kornblith 1993, 42). For example, referring to the essence of water as $H_2O$ provides a solid microstructural basis for explaining the observable properties of water in various circumstances.

In sum, since the 1970s it has become increasingly common to think that (1) at least some of our scientific concepts are natural kind terms, and that (2) they are based on the mind-independent causal structures in the world instead of only observed regularities. Furthermore, (3) kinds are assumed to be investigable by empirical methods (Boyd 1999). In the next section, I show how the mechanistic theory of natural kinds satisfies these sensible realist demands, but still applies to the often fleeting phenomena in the social and behavioral sciences.

## 3. Mechanistic theory of natural kinds

As was noted above, paradigmatic examples of natural kinds often come from the natural sciences: the kindhood of elementary particles or chemical substances is guaranteed by the shared microstructure of the instances of the kinds. However, consider some of the phenomena in the social and behavioral sciences discussed in the articles below. Human cognitive abilities such as memory or the capacity for conceptual thought are outcomes of the functioning of evolutionary processes, and they do not appear to be supported by universal and excep-

tionless laws of nature. Psychiatric phenomena like Asperger's, pathological gambling, or bulimia nervosa are deeply embedded in the surrounding social structures, and arguably the existence of culture-bound disorders depends on a specific social context. These kinds of phenomena are not only supported by the intrinsic properties of the members of the kind, but they also rely on the functioning of stabilizing mechanisms in the environment.

However, these phenomena are far from being stipulations or "mere" social constructions. Although they do not have intrinsic essences and they are not directly supported by laws of nature, they are stable and predictable enough to plausibly be considered as building blocks of our cognitive architecture and social reality. Moreover, as in the natural sciences, the kind concepts play a role in scientific explanations of human behavior and in designing interventions on phenomena.

As I mentioned in section 1, I use the phrase 'social and behavioral sciences' to refer to the constellation of scientific disciplines that all primarily aim to provide understanding of human behavior and its causes. In addition, my usage of the label also reflects my commitment to the belief that there is no deep qualitative dividing line between "harder" behavioral sciences, like biological psychology or cognitive science, and "softer" disciplines like sociology and psychiatry. Despite the theoretical quietism or even instrumentalist attitude of some scientists, research practices in several disciplines in the social and behavioral sciences suggest that classificatory decisions made in research reflect scientists' understanding of what the real phenomena studied in the discipline are. Therefore, a working hypothesis in the following articles has been that in the whole range of the social and behavioral sciences, despite the differences in their particular methods and styles of theorizing, the ultimate goal of research is the same: uncovering genuine phenomena (Bogen & Woodward 1988) and causal structures underlying them.

I hold that this methodologically monist position together with my epistemology-oriented interpretation of natural kindhood introduced in the previous section legitimatize my approach that treats research also in the social and behavioral sciences as being after natural kinds.[9] However, due to the reasons given above, traditional theories of natural kinds building on the notions of law and essence do not seem adequate for describing phenomena in the special sciences. The theory of classification in the social and behavioral sciences developed in the following articles builds on a more promising account of special science kinds, Richard Boyd's theory of *natural kinds as homeostatic property clusters*.

### 3.1. Natural kinds as homeostatic property clusters

In a series of publications spanning several decades, Richard Boyd (1980; 1983; 1991; 1999; 2010) has developed his naturalistic and anti-foundationalist version of scientific realism that tries to take into account insights both from constructionism and sophisticated forms of empiricism (Kuhn 1996; van Fraassen 1980), while still arguing for an undiluted form of scientific realism. His theory of natural kinds plays a central role in this endeavor. In the following, I introduce three central features of Boyd's theory that make it a promising starting point for my account of classification and concept formation in the social and behavioral sciences. First, Boyd understands natural kind terms as mediators between the inferential needs of a scientific discipline and the often complex causal structure of the world. Second, his approach is naturalistic in the sense that it implies that kind membership as such does not ultimately do any explanatory work; the use of kinds in scien-

[9] I should emphasize that I do not think that this methodological monism implies anything that could be called methodical monism. Different fields studying human behavior employ widely different methods, and the weak kind of methodological monism that I advocate only implies that the *ultimate epistemic and practical aims* of research across the social and behavioral sciences are the same.

tific inference is justified by empirical knowledge of the causal structures underlying a kind. The third central feature of Boyd's theory is that natural kinds need not be supported by universal and exceptionless laws of nature. Instead, they can be sustained by mechanisms that have limited scope and that are sensitive to changes in background conditions.

It could be said that Boyd's approach to scientific concepts has its roots in the issues raised by the grue paradox discussed in section 2.3. Boyd recognizes the role of conventionalist elements in concept formation, and characterizes conceptual change in science as a deeply theory-laden project, in which new theories and concepts are created based on new empirical findings, but also on existing theoretical knowledge (Boyd 1980; 1983). Describing concept formation as a dialectical process in which we iteratively devise more accurate and instrumentally reliable representations of targets in reality, Boyd characterizes the epistemic role of natural kind concepts as one of *accommodating* between the epistemic aims of a scientific discipline and the causal structure of reality (Boyd 1999, 56). I take it that the slightly opaque metaphor of accommodation is meant to convey the idea that, in order to count as a natural kind term, a concept need not describe the target of research perfectly accurately, or in full detail. Instead, it is sufficient that a natural kind term captures relevant causal structures of the target in a way that satisfies the inferential needs of the discipline in question.

In this sense, Boyd's theory sits firmly within the epistemology-oriented tradition of natural kinds: Boyd conceives of natural kind terms as being embedded in the epistemic practices of scientific inquiry, and his theory of natural kinds offers an explanation of where the inferential and practical reliability of these concepts comes from.

While Boyd's overall defense of scientific realism is intricate, the central idea of his theory of natural kinds is easy to understand. Accord-

ing to Boyd, many scientifically interesting natural kinds are homeostatic property clusters, and they consist of two elements:

{HPC}
α) a group of often co-occurring properties, and
β) a homeostatic causal mechanism that brings about their co-occurrence
(Boyd 1999, 67).

The idea is perhaps easiest to grasp through an example. Boyd's paradigmatic example of an HPC kind is a biological species.[10] Despite the variation within the population, the members of a species share a number of morphological, physiological, and behavioral properties because of the exchange of genetic material. The homeostatic mechanism supporting the cluster of shared properties is the interbreeding of conspecific populations combined with reproductive isolation from contraspecific ones – members of a species mate only with each other, not with individuals outside their species (ibid.). In subsequent discussions in the philosophy of science, this approach to the unity of kinds has been applied also to phenomena in the psychological and social sciences (Griffiths 1998; Machery 2005; Mallon 2003; Samuels 2009).

There are a few different interpretations of the notions of homeostasis and mechanism in this picture. On the one hand, Hilary Kornblith (1993, 35) understands mechanisms as the kind members' intrinsic properties that tend to occur together and account for the observable properties of the kind (Murphy 2006, 339). On the other hand, Paul Griffiths (1998, 188) adopts a far more liberal stance, according to which *any* theoretical structure that accounts for the projectibility of a category counts as its essence. Furthermore, in her account inspired

[10] In the philosophy of biology, the question of whether biological species are natural kinds has been the topic of a long-standing controversy (cf. Wilson 1999).

by Boyd's theory, Ruth Millikan apparently understands the reference to mechanisms in an etiological sense: a mechanism is the series of events that has brought about the co-occurrence of the properties in the cluster (Millikan 1999; 2000).

The *mechanistic theory of natural kinds* examined in this dissertation differs from all of the interpretations above.[11] In contrast to Griffiths, my version of the mechanistic theory suggests that not just any theoretical explanation for clustering will do, but instead, homeostatic mechanisms must be causal structures. However, in contrast to Kornblith, I do not think that all mechanisms are intrinsic to the members of the kind. Moreover, unlike Millikan, I understand the role of mechanisms in sustaining the property cluster in a *synchronic* manner. It is not the past but the current functioning of the homeostatic mechanism that explains the way that the properties cluster.

Therefore, the mechanistic theory is basically a two-level picture of natural kinds. Natural kinds are relevant from the point of view of inductive inference, because they have robust clusters of relevant properties that can be generalized to new instances of the kind. Knowledge of the homeostatic mechanism is useful for inductive inference in an indirect manner: knowing the mechanism can be used to *explain* the co-occurrence of the properties in the cluster, and the mechanistic information facilitates counterfactual inferences of how the cluster would change if the underlying causal structure or the background conditions were different.

Assembling a theory of concept formation around the notion of mechanism has some interesting advantages. In relying on causal in-

[11] My version of the mechanistic theory resembles those put forward by Machery (2009), Murphy (2006), Kendler, Zachar, and Craver (2010), Samuels (2009); and Wilson, Barker, and Brigandt (2007). I gratefully acknowledge some intellectual debts to these accounts. However, as becomes evident in the articles below, especially my views on the nature of mechanisms and on classificatory pluralism set my approach apart from many of these positions.

formation as the foundation for classification, the mechanistic theory obeys the naturalistic order of explanation: Delineation and identification of kinds, and projectibility judgments regarding their properties, are made *a posteriori*, based on empirical findings about the targets of research. Ultimately kind-membership as such does not have explanatory force, but when kind terms are employed in scientific explanation, claims about projectibility of predicates are justified by knowledge of the homeostatic mechanism of the kind. The mechanistic theory thus fulfills Quine's (1969, 52) requirement that mature fields in science should base their inductive inferences on knowledge of the structures underlying kind-dispositions, not on unexplained similarities between the members of kinds.

Replacing the notion of law with that of causal mechanism also makes the account a promising contender for a theory of concept formation in the special sciences, where exceptionless laws of nature cannot be found. Finally, employing the notion of mechanism in a theory of classification suggests a way to bridge two separate discussions in the philosophy of science: Although there is no satisfactory analysis of the notion of mechanism in the literature on natural kinds, it is possible to rely on the extensive discussions on the concept in the philosophy of biology and the philosophy of cognitive sciences to elaborate this central notion of Boyd's theory (see section 4).

### 3.2. Mechanistic theory and conceptual change

A large part of the appeal of the mechanistic interpretation of Boyd's theory of natural kinds has stemmed from its usefulness as a foundation for a theory of conceptual change in the life sciences. Based on the theory, it has been suggested that familiar notions such as EMOTION, MEMORY, or CONCEPT do not form genuine natural kinds, and should therefore be eliminated from the scientific vocabulary (Craver 2004; Griffiths 1998; Machery 2009).

The central idea of the accounts of conceptual change building on Boyd's theory is that we should align our concepts with mechanistic

structures in reality. In brief, what I call the *split-lump-eliminate scheme* (SLE) is based on three operations [see VI]. Firstly, if a concept refers to several different mechanisms, we should split it so that each mechanism gets its own corresponding concept. Secondly, however, a concept should capture the maximal class of phenomena sustained by the same mechanism. Therefore, if it turns out that a group of phenomena that were previously considered as separate are supported by a common mechanism, we should lump these phenomena under the same concept. And thirdly, if no well-defined mechanism corresponds to a natural kind concept, the notion should be eliminated from scientific vocabulary. The core idea underlying these three operations is that there should be a one-to-one correspondence between scientific concepts and mechanisms in reality (ibid.).

The SLE scheme has been particularly popular in the philosophy of psychology. In comparison to its recent predecessor as a picture of conceptual change in this domain, eliminative materialism, the SLE scheme is an advancement in several respects. It offers a more fine-grained picture of conceptual revision by not only focusing on elimination but also including cases of unification and non-eliminative conceptual refinement as species of conceptual change. Moreover, unlike many of the earlier discussions, the model does not rely on semantic intuitions about reference as a basis for conceptual change (Stich 1996).

### 3.3. Theories of concepts and kinds. Sketching the roadmap

I now briefly recapitulate the main points made in the past two sections. In section 2, I suggested that psychological and semantic theories of concepts should be thought of as two different perspectives on conceptuality. While a theory of scientific concepts – like the mechanistic theory examined here – should be compatible with insights from both these perspectives, it is not entirely clear how substantial constraints the psychological concept theories and the discussions on concepts in the philosophy of mind and the philosophy of psychology

could impose on a theory of scientific concepts: As Boyd's theory suggests, scientific concepts are typically embedded in networks of theoretical inference and revised based on knowledge achieved by theory-laden methods. Consequently, it is not clear how contributions from either psychological research on simple perceptual concepts, or, for instance, information-theoretic accounts of representation in the philosophy of mind (Dretske 1981; Fodor 1998) should be connected to the current perspective.

In addition, the theories of natural kinds comprise a motley crowd. Discussions on natural kinds serve several distinct conceptual aims, and I suggest that the literature in the philosophy of science primarily motivated by the inferential reliability of scientific concepts is somewhat distinct from theories of natural kinds in philosophy of language and metaphysics [I]. I should emphasize that, obviously, there are interrelations between the different ways of theorizing about kinds. My aim here has been merely to shift the burden of proof so that the shared aims and presuppositions of different theories cannot just be assumed, but they must be argued for.

Although the distinctions made in the sections above inevitably remain tentative, I believe that they are useful, especially because many theories of (scientific) concepts and natural kinds in the existing literature do not explicitly position themselves in relation to other theories. This situation easily creates confusion, or at least, it often remains unclear which theories one should pay attention to in one's own theorizing about concepts.

The task of the two previous sections has been largely negative in spirit. By suggesting that philosophical discussions on concepts and kinds do not form a unified debate, I have tried to free up conceptual space for developing my theory of scientific concepts. I next turn to what I believe is a more fruitful source of intellectual resources for elaborating a theory of conceptual change in the social and behavioral sciences – philosophical discussions of scientific explanation and

mechanistic research heuristics. Perhaps the most crucial conceptual maneuver performed in the articles in this dissertation is to develop Boyd's theory by connecting it to insights from the recent philosophical literature on the nature of mechanisms and mechanistic explanation. Such work is helpful, because as commentators of Boyd's theory have pointed out, the notion of mechanism employed in the theory is not altogether clear (Craver 2009; Reydon 2009). I argue that closer examination of the notion in Boyd's theory results in a more developed picture of how kinds and mechanisms are identified, and where their inferential usefulness derives from.

## 4. Explanation and mechanisms

The existence of a close relationship between classification and explanation has been recognized both in the psychological and philosophical literature (Gopnik & Meltzoff 1997; Griffiths 1998; Lombrozo 2009). However, often treatments of the topic have not relied on a systematic theory of explanation, and I believe that the fruits of applying the latest developments in the theory of explanation to questions concerning the foundations of classification have not yet been picked.

In this section, I introduce the theory that I regard as the most promising contender for an account of causal explanation in the social and behavioral sciences, *the contrastive-counterfactual theory of explanation* (CC-theory) (Woodward 2003; Ylikoski 2001). I also review parts of the recent philosophical literature on mechanisms. Although the mechanistic approach to explanation and the contrastive-counterfactual theory have sometimes been seen as competing pictures of explanation (Waskan 2011), I think that it is also possible to think of them as being complementary (Craver 2007; Glennan 2005; Steel 2008; Ylikoski 2011). In my view, the division of labor between the theories is roughly as follows: the mechanistic approach to explanation offers a realistic view of research practices in the sciences, and CC-theory can be used to elaborate its conceptual foundations. Firstly, CC-theory offers a non-reductivist account of mechanisms suitable

for describing the biological, cognitive and social mechanisms occurring in the social and behavioral sciences. It also, secondly, provides a systematic criterion for demarcating the boundaries of mechanisms and, thirdly, it suggests that *prima facie* competing mechanistic explanations phrased at different levels of description might often turn out to be complementary. Hence, it offers useful resources for thinking about interfield cooperation in scientific research.

### 4.1. Contrastive-counterfactual theory of explanation

A shared starting point for many theories of scientific explanation has been to distinguish explanation from other epistemic activities (e.g., description and prediction) by pointing out that explanations offer information of a specific kind: explanations tell *why* or *how* something happened. According to the still most widely-known model of scientific explanation, the deductive-nomological model (DN-model), explanation is essentially a matter of subsuming explanandum-events or higher-level laws under the laws of nature.[12] The model articulates the notion of explanatoriness as nomic expectability; according to Hempel (1965) we have explained a result when we could predict it as an outcome of the laws of nature and the relevant initial conditions.

In the following articles, I rely on what Petri Ylikoski and Jaakko Kuorikoski call the *contrastive-counterfactual theory of explanation* (Ylikoski 2001; Kuorikoski 2010). This theory is largely similar to James Woodward's (2003) well-known account of causal explanation, and my view is also indebted to Woodward's theory. Compared to the DN-model, the contrastive-counterfactual theory takes a clearly different approach to explanation. It suggests that explanatory why- and how-questions are answered by *tracking change-relating counterfactual dependencies* between the relata in the explanation. Successful explanations uncover objective relations of counterfactual dependency be-

[12] For good summaries of the development of theories of explanation in the philosophy of science, see Salmon 1990 and Ruben 1990.

tween things in the world. Roughly, in the case of causal explanation, had the cause variable been intervened upon, the effect would have been different.

A second fundamental principle of CC-theory is that, at least implicitly, explanations have a contrastive focus (Garfinkel 1981; van Fraassen 1980; Woodward 2003; Ylikoski 2007). They can usually be characterized as answers to questions of the form: why fact rather than [foil], where the foil is an exclusive alternative to the fact. Explanatory knowledge thus has the following form:

| | |
|---|---|
| {EK} *x [x'] because of y [y']* | (variable X takes the value x instead of x' because Y has the value y instead of y') |

This characterization of explanatory knowledge allows us to clearly specify the difference between explanatory and descriptive knowledge: Explanations allow us to answer counterfactual *what-if-things-had-been-different questions* (Woodward 2003). That is, while descriptive knowledge provides answers to *what, when, where*, and *how-much* questions regarding the factual properties of phenomena, explanatory knowledge of invariant relationships tells how changes in *explanantia* (explanatory factors) would influence *explananda* (things to be explained) (Ylikoski 2011).

Such an approach to explanatory knowledge breaks the problematic symmetry between prediction and explanation assumed by the DN-model. CC-theory implies that despite there being problems with predicting the molar behavior of complex systems, tracking dependencies between variables offers a way to uncover and understand causal processes underlying the systems in a piecemeal way. This distinction is useful because it can be used to reply to antinaturalist arguments that cite the impossibility of accurate prediction in the social sciences as a reason for rejecting explanation as their theoretical aim.

Moreover, by not making reference to laws of nature, CC-theory avoids another antinaturalist argument. The sharp division between natural and social sciences has been defended by arguing that phenomena in the social sciences are not amenable to nomological explanation, and indeed, law-based accounts are widely held to be problematic pictures of explanation both in the psychological and the social sciences (Cummins 1983; Craver 2007, ch. 2; Hedström & Ylikoski 2010). However, as CC-theory conceives of explanation as tracking dependency-relations between variables, it avoids many of the problems of the law-based accounts. Explanatory dependencies need not be universal or exceptionless, only a modest degree of invariance is needed. CC-theory also offers conceptual resources for precise assessment of the scope and stability of these explanatory relationships, and consequently, it can be used to characterize the often local and exception-ridden nature of special science generalizations in a more analytical way than the one offered by *ceteris paribus* accounts of special science laws (Woodward 2008; Ylikoski & Kuorikoski 2010).

## 4.2. Explaining with mechanisms

In many parts of the life sciences, it is common to think that a genuine explanation of a phenomenon requires providing a description of a mechanism (Wright & Bechtel 2007; Ylikoski 2012). The concept of mechanism has recently acquired an important role also in the philosophy of science literature. These discussions have mainly focused on the ontology of mechanisms and the search for mechanisms as research heuristics in the life sciences.

According to traditional intuitions about mechanisms, relationships in mechanistic structures are based on solidity, rigidity, and impenetrability, and the explanatory usefulness of mechanism description derives from these properties. In a similar manner, Wesley Salmon's (1984; 1998) causal-process account of explanation relies on the spatio-temporal continuity of mechanisms, and their ability to transmit conserved quantities such as energy or momentum. However, in many

recent discussions in the life sciences, a broader notion of mechanism has been employed. Since the seminal paper by Machamer, Darden, and Craver (2000), mechanisms have often ontologically been understood as causal structures consisting of entities and their activities. In different disciplines, the entities and activities can be of different complexities, and for instance in the cognitive sciences, the activities can often be described in information-processing terms (Bechtel 2008). In sum, it has become common to understand a mechanism in broad sense as (i) a collection of causal parts (ii) organized together to sustain a stable (iii) phenomenon (Bechtel & Abrahamsen 2005; Glennan 2002; Woodward 2002).

In their work, William Bechtel and Robert Richardson (2010) have studied the role of mechanistic research heuristics in the practices of scientific explanation. In the life sciences, a typical explanatory task is to understand how a complex capacity of a system is made possible by its material structure. Bechtel and Richardson characterize the reductivist research heuristics employed in attacking such problems as consisting of three stages. First, the explanandum phenomenon is decomposed into a set of simpler functional units. Then, the system in question is structurally decomposed into its components parts. Finally, one tries to fit these decompositions together by localizing the functional capacities on the structural components of the system.

Together, the three stages amount to a strategy for reverse-engineering complex systems: constitutive explanation by functional decomposition and localization produces hierarchical multi-level descriptions of systems, where a higher-level explanandum capacity ($\psi$-ing) is explained in terms of the organized functioning of lower-level capacities ($\varphi_1, \ldots, \varphi_n$), which in turn are explained by decomposing them into yet lower level capacities ($\rho_1, \ldots, \rho_m$) (Figure 1). By showing how the material parts of the system and their organization conspire to constitute the capacities mentioned in the functional decomposition of the explanandum phenomenon, the explanation discloses

how the explanandum is brought about by a multi-level constitutive mechanism.[13]

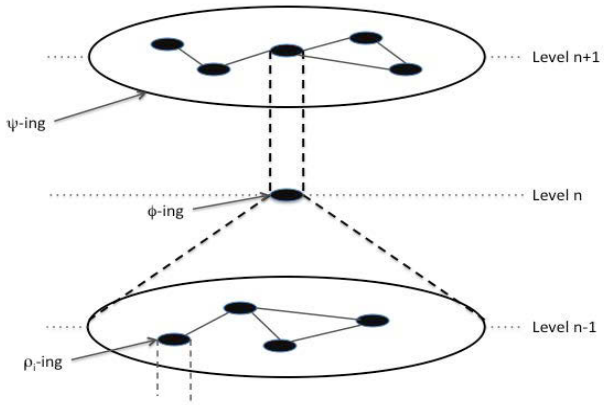


**Figure 1. Craver diagram (cf. Craver 2007, ch.5)**

## 4.3. Contrastive-counterfactual analysis of mechanistic explanations

I now suggest that CC-theory can be seen as giving a systematic account of why describing mechanisms is explanatory in the first place. The analysis of the nature of explanatory knowledge in section 4.1. suggests that describing mechanisms allows us to answer *what-if-questions* regarding what would happen to the explanandum phenome-

[13] As shown by Ylikoski (2012), there are important differences between causal and constitutive explanation. Whereas causation is a relation between distinct events and it takes time, constitutive relationships hold between the properties of the whole and its parts, and do not manifest the manipulational asymmetry characteristic of causal relations. Despite the differences between causal and constitutive explanation, both can be understood as tracking dependencies between explanantia and explananda.

non, if there were changes to the components of the mechanism or their organization. Describing mechanisms allows us to answer how-questions by "opening black boxes," by revealing more detailed descriptions of the processes connecting causes and effects. Mechanistic explanation is therefore a species of reductive explanation in the weak sense that it aims to provide understanding of the properties of the whole in terms of the properties and organization of its components.

However, CC-theory makes it explicit that constitutive mechanistic explanations need not amount to strong reductivism. Such explanations do not explain away higher-level explananda, but only articulate how they depend on their parts. CC-theory suggests that mechanisms can be found at several levels of analysis, if it is possible to analyze occurrences at a level in terms of the functioning of components, which are modular in the sense that it is possible to change the behavior of one part independently of the others, and if their behavior conforms to generalizations that are invariant at least under some interventions (Woodward 2012). Thus, the contrastive-counterfactual theory offers a *non-fundamentalist account of mechanisms* suitable for describing the biological, cognitive, and social mechanisms occurring in the social and behavioral sciences.

Another advantage of amalgamating CC-theory with mechanistic theories of explanation is that CC-theory provides a systematic criterion for mechanism identification and for demarcating their boundaries. It is generally agreed that a good explanation has to distinguish between relevant explanatory factors and mere causal background conditions, but earlier accounts of mechanistic explanation, such as Salmon's causal-mechanical theory, typically lack a satisfactory analysis of explanatory relevance (Hitchcock 1995). By contrast, CC-theory provides a relevance criterion for mechanism demarcation: as I have argued, the theory suggests that, given a cluster of explananda (corresponding to the explanatory aims at hand), an explanatory mechanism should include those, and only those, factors that are needed to explain the contrasts between explanandum states and their contrastive

foils [IV]. This suggests that a mechanism could be seen as a chunk of the causal structure of reality, isolated for explanatory purposes.

For similar reasons, CC-theory provides a workable starting point for analyzing relationships between different mechanistic explanations of the same target phenomenon. The contrastive nature of explanatory knowledge suggests that explanations are always aspectual, and that explanations do not explain phenomena as such, but only answer certain contrastive questions about their properties. Consequently, often *prima facie* competing explanations of a phenomenon might turn out being compatible, once their different contrasts are properly spelled out. For example, laboratory and ecological approaches in the psychological study of human memory often pick their variables of interest differently, and delineate the studied systems in different ways. They can answer different contrastive questions about the target, and need not necessarily be seen as mutually exclusive approaches to studying the target (see section 6.4.).

## 4.4. Setting the agenda

The general research agenda in the following articles arises from the theoretical background introduced in the past three sections. I hold that the homeostatic property cluster theory of kinds together with the literature on scientific explanation and mechanisms provide a promising starting point for a naturalistic theory of conceptual change in the social and behavioral sciences. However, several questions remain regarding the applicability of the mechanistic approach in the social and behavioral sciences.

(1) Both Boyd's theory of kinds as well as the theories of mechanisms originate in discussions in the philosophy of biology, and extending them to the social and behavioral sciences raises novel conceptual problems. In order to pass as a satisfactory description of research practices in the social sciences, it needs to be shown that the mechanistic model can be employed also as an explanatory model of intentional and intersubjective action. In particular, as is suggested in

the articles below, complex criss-crossing causation between factors from social, personal-level, and subpersonal spheres creates considerable challenges to mechanistic explanation in the social and behavioral sciences (Hacking 1995b).

(2) The contrastive approach to the analysis of explanatory relevance based on CC-theory suggests that the way that mechanisms are identified and demarcated, and how they should be described, are relative to the explanatory aims at hand. As argued by Craver (2009), this seems to imply that the mechanistic theory of natural kinds fails to provide unambiguous judgments of where the kinds are, and different explanatory aims can lead to different ways of classifying the same domain of phenomena. This seems to challenge the usefulness of the mechanistic theory as a normative account of scientific classification.

(3) Several indispensable scientific concepts raise concerns about the scope of the mechanistic theory of kinds and the SLE scheme of conceptual change: many concepts (e.g., GENE, RATIONALITY, INFORMATION) do not appear to correspond to a single well-defined mechanism, but they still play an important role in scientific research. Therefore, whether the mechanism-based account can provide an all-encompassing picture of scientific concept formation is an issue that requires further study.

## 5. Overview of the articles

Because the papers are an outcome of several years of work revolving around the same set of issues and attacking the same problems from various angles, they are connected to each other by messy relations of family resemblance. Between each pair of articles, it is possible to identify several common threads. In consequence, deciding how to put them in a linear order was not easy. I have decided to organize the articles in roughly the order that they were written, as this ordering reflects the development of my thinking on the topic, and also corresponds to the three open questions to the mechanistic approach introduced at the end of the previous section: The six articles can be

divided into three pairs. The first two examine how the mechanistic theory of natural kinds can be applied to psychiatry and the social sciences. The second pair delves deeper into the conceptual machinery that I utilize by examining the central notion of mechanism, and discusses questions of mechanism description and identification in the mind sciences. The last two papers examine challenges to the mechanistic theory raised by non-modular target systems and non-mechanistic scientific concepts.

## 5.1. Carving the mind by its joints: Culture-bound psychiatric disorders as natural kinds

*Carving the mind by its joints* provides a common starting point for the rest of the articles by introducing the mechanistic theory of natural kinds and by distinguishing this notion of natural kind from other uses of the concept in other parts of philosophy. The paper picks up a problematic culture-bound psychiatric disorder, bulimia nervosa, and argues that such problematic phenomena sustained by both physiological and social causal factors should be understood as natural kinds, instead of as social constructions.

Argumentatively, the paper is structured around a dilemma. On the one hand, bulimia appears as a classic contrast case to natural kinds. The eating disorder is sustained by a hybrid set of causal factors. While research suggests that there are physiological and genetic causes that make certain individuals susceptible to bulimia, the illness is also sustained by norms regarding slimness and beauty, and is thus dependent on social factors. Therefore, bulimia does not appear to count as a natural kind: it is not supported by a law of nature, it requires a specific socio-cultural niche to exist, and in consequence, it is spatiotemporally local. On the other hand, calling it a "mere" social construction would be misleading as well. There are robust generalizations to be made about the illness, it is sustained by real – albeit partly social – causal structures, and the psychological and societal consequences of the phenomenon are evident.

My solution to the dilemma proceeds by conceptual analysis of the notion of natural kind. I argue that the tradition of natural kinds consists of several partly independent strands, and the notion of natural kind is used in importantly different ways in the philosophy of language, metaphysics, and the philosophy of science. The central claim of the article is that although in the light of discussions in, say, metaphysics, calling culture-bound phenomena such as bulimia natural kinds would appear odd, for the purposes of the epistemology-oriented tradition of natural kinds, such a strategy is coherent. Since Mill (2002), the central motivation in the epistemological tradition has been to distinguish scientifically pertinent phenomena and concepts from conventional or erroneous ones, and to thereby identify concepts that can reliably support our epistemic practices (see section 2.3. above). My solution thus slips through the horns of the dilemma by making more precise the notion of natural kind which is relevant to the concerns motivating discussions on natural kinds in the philosophy of science.

The second goal of the paper is to argue that the mechanistic theory of natural kinds should be interpreted as a contribution to this epistemology-oriented tradition of natural kinds, and that the theory offers a promising platform for approaching complex interdisciplinary targets such as bulimia. I show how within the mechanistic framework, culture-bound phenomena can be conceived of as natural kinds that are sustained by *hybrid mechanisms* comprising of neural, psychological, and social submechanisms.

Extending the theory of natural kinds to (partly) socially supported phenomena in the social and behavioral sciences has three main epistemic advantages. As pointed out by Hacking (1999), the social constructionist rhetoric is often unclear whereas, by contrast, the mechanistic theory (1) offers a more analytic approach to understanding sociological causation. The reliance on counterfactual mechanistic information (2) provides conceptual tools for assessing the reactions to treatment interventions in different contexts. And the hybrid-

mechanisms approach (3) provides a promising framework for conceptual and explanatory integration between different fields working on the same phenomenon. In psychiatry, such an integrative approach could also have considerable practical value by downplaying the stark controversies between supporters of pharmacological and therapy-oriented approaches to treatment.

## 5.2. Looping kinds and social mechanisms

Classifying people into categories has always been a strategically important task for the social sciences. Classifications make it possible to localize social problems into certain groups of people and makes them easier to control and prevent. However, it has been widely recognized among social scientists that the practices of classification themselves have important social consequences (Bourdieu & Wacquant 1992; Giddens 1984), and studies of how labeling and classification influence the behavior of the classified have played an important role in fields such as criminology, sociology of deviance, mental illness, political sociology, and sociology of science. Given the increasing cultural visibility of neuroscience and genetics, this relationship between scientific knowledge and behavior is as important as ever. Knowledge of one's genetic susceptibility to certain illnesses or personality traits, as well as the novel neuro-discourse, can be rich sources of self-conceptualization for the modern individual (Singh & Rose 2009).

The capacity of labels and classifications to influence identity and behavior appears to be based on three features of classifications of humans: (1) they are usually dependent on socio-cultural practices and often local in space and time; (2) medical and social-scientific classifications of humans are normatively loaded; and, (3) their targets are conscious agents. In his now classic set of case studies, Ian Hacking (1995a, 1998) has studied the *looping effect* of human kinds, which is a consequence of these features of classifications in the social and behavioral sciences. The dynamics of looping can be briefly described in the following way: new classifications of people often provoke reac-

tions in those classified, and this creates a need to revise the original classification to fit the new property profile of the target group. The new knowledge might again provoke unpredicted reactions in targets, and this feedback-connection between behavior and scientific knowledge sets the looping in motion.

The looping effect can seem like a serious obstacle to the application of the mechanistic theory of kind-concepts in the social and behavioral sciences. The reflexivity typical of looping kinds has motivated claims of social construction of the phenomena in question. More generally, looping can be interpreted as an instance of the kind of sensitivity of human phenomena to meanings and intentions that sets the social and behavioral sciences methodologically apart from sciences employing causal-mechanistic explanation. Moreover, looping makes human scientific phenomena volatile, and this has often been seen as a general challenge to the realistic understanding of concepts and classifications in the social and behavioral sciences.

This article continues the line of argument begun in the first paper. Jaakko Kuorikoski and I argue that it is possible to incorporate looping human kinds within the mechanistic model of natural kinds. Furthermore, we propose that the kind of social construction implied by looping is not an insurmountable challenge to a causally realist approach to classification in the social and behavioral sciences. By relying on a set of examples drawn from social psychology, economics, and psychiatry, we illustrate how looping effects can be described in terms of the functioning of a set of different kinds of social and cognitive *causal feedback mechanisms* that mediate the effect. Furthermore, we argue that the fleeting cases of looping kinds described in Hacking's case studies capture only a special form of a more general phenomenon, there being also situations in which interactivity stabilizes phenomena. In fact, we suggest that mechanisms similar those behind the looping effect often sustain relatively stable social institutions. Often the seemingly spontaneous order of social reality is founded

upon shared and self-reinforcing expectations concerning appropriate behavior.

The mechanistic picture of looping kinds offers several advantages compared to Hacking's original account, which portrays looping in "semantic" terms, as a relationship between action and meaning (Murphy 2006, ch. 7). Most importantly, by stressing the importance of having knowledge of the causal processes underlying reflexivity, it accounts for the fact that people's awareness and interpretations of classifications only become social reality when they are acted out. Looping is sustained by social and cognitive mechanisms, and knowing these mechanisms allows one to answer a large set of counterfactual *what-if* questions that merely descriptive accounts of looping effects fail to answer.

## 5.3. Intentional concepts in cognitive neuroscience

There has been an extensive discussion on mechanisms in the recent philosophy of science (Bechtel & Abrahamsen 2005; Glennan 2002; Machamer, Darden & Craver 2000). However, the literature has mainly focused on ontological questions concerning the nature of mechanisms and their parts, and on epistemic considerations regarding the discovery of mechanisms and their explanatory role. The topic of this paper, the question of how mechanisms ought to be (linguistically) described has received considerably less attention.

In their controversial work, Maxwell Bennett and Peter Hacker (2003; 2008; 2009) (B&H) have laid out a wealth of examples from psychological and neuroscience research showing that in the mind sciences it is common to describe the functioning of the brain and its parts with intentional predicates. The famous claim of these authors is that in describing neural and cognitive structures in intentional terms – as believing something, making inferences, or forming hypotheses, to mention just a few examples – neuroscientists make a conceptual mistake that nullifies the explanatory power of their theories. In attributing cognitive capacities belonging to the whole (the human

agent) to its parts (the brain), neuroscientists are said to commit the *mereological fallacy*, resulting in the creation of nonsense.

Although B&H's critique is exaggerated, and often relies on outdated notions of explanation, and on a rigid view of functioning of common sense psychological concepts, neither do I find the most prominent replies to their critique satisfactory (Churchland 2005; Dennett 2007). In stark contrast to B&H, their adversaries like Daniel Dennett have often adopted an attitude of wholesale acceptance towards the contested language use. I show that in cognitive neuroscience, there are both successful and somewhat confused uses of intentional predicates in mechanistic explanations, and subsequently, it should be considered a criterion of adequacy for a theory of subpersonal intentional language use that it can differentiate between these two types of cases.

My solution to the problem raised by the use of intentional predicates in neuroscience starts from a switch of perspective from semantic to explanatory: to understand the contested language use, one must pay attention to its explanatory context. Following the heuristics of decomposition and localization approach to psychological explanation (see section 4.2.), I suggest that the relationship between mental phenomena and their neural basis should not be understood as a mapping between two levels, the mental and the physical (Kim 1992), but instead as a hierarchy of abstract descriptions of the same system (Bechtel & Richardson 2010; Craver 2007). This view suggests that there is no sharp dividing line between levels that are acceptably described in representational terms and "purely" mechanistic ones (Lycan 1990). However, there are important ways in which the mechanistic account of explanation constrains the use of common-sense psychological concepts in subpersonal explanations. In brief, I suggest that (1) inferential commitments made in mechanism descriptions must be compatible with lower-level descriptions of the same material structures and (2) the explanatory import of functional decompositions depends on mechanistic parts being simpler than explananda.

Working out the consequences of these normative principles unfolds my account and my criticism of both B&H and Dennett's intentional-stance approach.

My own position could be called cautious liberalism about mechanism description. I propose that inferentially rich common-sense psychological concepts (guessing, inferring, etc.) can in principle be used to describe subpersonal levels, if such concepts adequately capture functional roles and operations at that particular level of inquiry. Moreover, by means of an example drawn from vision research, I suggest that *intentional predicates can play important heuristic role in mechanistic explanations as tools for functional decomposition*, and work as a necessary theoretical scaffolding on the way to more precise mechanistic and computational explanations. This suggests a dynamic picture of mechanism description, in which looser terms are replaced by more precise ones as research progresses.

Finally, in the last section of the paper, I show that often the general suspicion against mechanistic explanation of intentional-level phenomena relies on mistaken presuppositions about the nature of constitutive explanation. Opponents of mechanistic explanation of the mind have often pointed out that personal-level phenomena have properties (e.g., normativity, rationality) that do not have counterparts at subpersonal levels, and argued that this makes intentional phenomena irreducible to subpersonal goings-on (Hornsby 2000; McDowell 1996). To answer this challenge, I emphasize the non-reductivist nature of constitutive explanation. Mechanistic constitutive explanations do not amount to conceptual identifications between personal-level phenomena and subpersonal processes; explaining does not mean explaining away, but instead it provides answers to what-if questions regarding the relationships between the properties of the whole and the properties of the component parts. This kind of explanandaquestions do not exhaust the intentional phenomenon under study: in order to achieve comprehensive explanations of social and intention-

al-level phenomena, reductive heuristics must be combined with non-reductivist research methods.

## 5.4. Explanatory power of extended cognition

Whether human cognition is partly constituted by components residing outside the skull is the question at the heart of an extensive debate in the philosophy of psychology. In *Explanatory power of extended cognition*, I approach the topic of extended cognition as a question of scientific concept formation. I regard this perspective as well motivated but underrepresented in the contemporary philosophical debates on extended cognition. The early philosophical discussions on distributed and extended cognition were largely inspired by interesting developments in the psychological sciences, and many of the examples were drawn from scientific research (Clark 1997; Hutchins 1995; Hutchins 2010). However, as the philosophical debates on the topic have become perhaps more conceptually sophisticated, the intimate connection to scientific research has often been lost.

A common argumentative move in the debates has been to appeal to the explanatory power of the extended cognition hypothesis to assess its plausibility (Barker 2010; Clark 2007; Rupert 2004; Sprevak 2010). However, both proponents and critics have based their arguments on intuitive notions of explanatory power, and no systematic account of the notion has been offered in the debate. In this article, I try to fill this lacuna by applying CC-theory and the mechanistic theory of kinds to the problem of cognitive system demarcation.

Relying on the contrastive-counterfactual theory of explanation, I put forward my *differential influence criterion* for cognitive mechanism demarcation. In brief, I propose that a cognitive system ought to include only those variables, the variation in which explains contrasts in the cluster of explananda at hand. That is, while background conditions typically make a causal difference to the target of explanation, the variation in them is such that it cannot distinguish between the explanandum state and its contrastive foil. By reviewing a set of ex-

amples from the literature, I show that my criterion leads to reasonable judgments of cognitive system demarcation. It blocks cognitive bloat, but in principle allows for cognitive extension. That is, while differential influence does not regard the boundaries of the organismic skin-bag (Clark 2008, xxvii) as an obvious way to demarcate cognitive systems, it makes explicit why not all causal factors connected to a phenomenon should be treated as its genuine components.

A crucial implication of the differential influence approach is that cognitive system demarcation becomes dependent on explananda. This, in turn, leads to *cognitive systems pluralism*: in light of different explanatory aims, different scientific fields can end up with different demarcations of cognitive system boundaries. I suggest that given the complex nature of phenomena studied in the cognitive sciences, such classificatory pluralism is tenable, and that it can be understood as a consequence of the division of cognitive labor between research perspectives.

The differential influence criterion implies that once a set of explanandum-questions is held fixed, questions of explanatory relevance are settled and, consequently, so is the question of mechanism demarcation. In other words, given explananda, the differential influence criterion results in unambiguous judgments of cognitive extension. The principle in itself, however, offers no assistance in assessing the explanatory power of externalist explanatory strategies in comparison to intracranialist ones. Therefore, the differential influence criterion has to be supplemented with further analysis of the notion of explanatory power. By relying on Ylikoski and Kuorikoski's (2010) inferentialist account of explanatory power, I analyze an example case drawn from memory research. I argue that this approach to explanatory power that builds in the contrastive-counterfactual theory provides a powerful tool for constructing a view of the epistemic coordination between different research perspectives. It suggests that different classificatory strategies employed in laboratory research and in ecological approaches to human memory are characterized by different *pro-*

*files of explanatory virtues* and are thus fitting for answering different sets of explananda. In consequence, my approach draws a picture of cognitive system demarcation as one of finding an optimal fit between the cluster of explananda relevant to a scientific field and the corresponding explanatory mechanisms.

Regarding extended cognition as an ontological thesis, my conclusion in the paper is deflationary. Cognitive systems are cheap – systems and mechanisms can be demarcated in various ways, and looking for "the" correct cognitive kinds appears a misguided aim. However, comparing the explanatory power of different classification schemes suggests a way to move beyond inarticulate classificatory pluralism, and allows one to assess the characteristic epistemic strengths and weaknesses of extended and intracranialist approaches.

## 5.5. Understanding non-modular functionality – Lessons from genetic algorithms

The third pair of articles discusses challenges to the mechanistic model of scientific concepts. First of the two, *Understanding non-modular functionality – Lessons from genetic algorithms* concerns the limits of mechanistic explanation in the cognitive sciences. The paper elaborates the observation that evolutionary design processes can create tangled and complex causal systems that are hard to explain and understand, and since the human cognitive system is a product of biological evolution, it might resist our successful applications of the mechanistic research heuristics.

According to a common view, evolutionary designs are often opaque because they cannot accurately be analyzed by functional decompositions inspired by engineering-intuitions and rational reconstruction (Clark 1997; Marcus 2008). The problem with these kinds of observations has usually been that while they are not especially controversial, the exact nature of the problem has been hard to articulate, and it is still not clear how, precisely, evolutionary design creates barriers for understanding or mechanistic explanation. In this paper,

Jaakko Kuorikoski and I try to shed light on this problem by introducing a simple genre of computational processes that mimics evolutionary design, genetic algorithms. We suggest that evolutionary design processes create at least two kinds of problems for mechanistic understanding and for the heuristics of functional decomposition and localization. They lead to (1) psychologically foreign designs that often show (2) lack of both functional and structural modularity. Our main focus is on the second challenge, which has received less attention in the literature (Jacob 1977).

Bechtel and Richardson (2010) observe that a precondition for the successful application of the heuristics of decomposition and localization is that the target system is nearly decomposable (Simon 1962). We argue that a similar property of the architecture of complex systems, *modularity*, is needed for the functioning of these heuristics. The contrastive-counterfactual theory of explanation can make the rationale behind this requirement clear: constitutive explanations allow inferences regarding how the whole would change, were its components or their organization different. However, this requires that individual components can be intervened on without disturbing the functioning of other parts. In contrast, if it is not possible to produce a representation of the target system in terms of parts and variables that can be varied in a modular manner, changes to parts ramify uncontrollably and the inferential usefulness of the mechanistic model is reduced.

The issue of the modularity of the human cognitive architecture has played an important role in the history of the mind sciences (Fodor 1983; Uttal 2003; Zola-Morgan 1995). The question is of course an empirical one, and a definitive answer is still beyond the reach of our scientific knowledge. However, to get a firmer conceptual grip on the kind of problems that evolutionary design raises for mechanistic explanation, we mimicked some central structural features of evolutionary design processes in a computer simulation. We replicated a simple genetic algorithm (GA) simulation, where the GA was used

to develop a DNA for controlling the behavior of a simulated robot in a simple can-picking game. We then compared the solutions produced by the algorithm to ones written by a human designer (Mitchell 2011; cf. Goldberg 1989; Holland 1975; Mitchell 1998).

The simulation confirmed our expectations about the nature of evolved designs, and the solutions created by the GA clearly illustrate the problems to understanding raised by non-modular design. Successful solutions to the problem developed by the algorithm employ efficient behavioral strategies that cannot be deciphered simply by examining the isolated structural elements that determine the behavior of the simulated robot. Instead, looking holistically at the broad phenotypic behavior of the robot is necessary to detect the behavioral "hacks," which make the evolved solutions more efficient than the ones arising from engineering-minded design strategies. The problem with these highly efficient middle-level behavioral strategies is that while they are easy to visually detect in behavior, they are (1) holistically distributed across the DNA of the robot and have no dedicated structural basis. A change in a particular locus in the DNA often changes several behavioral strategies at once. Relatedly, the strategies are not functionally modular in the sense that they are not discrete but consist of collections of separate adaptations in different parts of the DNA. In consequence, these middle-level behavioral patterns are (2) not structurally localizable in any particular part of the DNA, but only trivially in the whole genome. Moreover, there are (3) strong interaction effects between the fitness-effects of the strategies and very specific selection environments. Often a strategy is adaptively valuable only in environments that are very similar to the ones in which it evolved. What is common to all these challenges is that they make it hard to derive robust counterfactual generalizations regarding the relationship between particular components of the system and its overall behavior, and thus they reduce the usefulness of mechanistic research heuristics.

The example discussed in the article is of course highly artificial, and the important dissimilarities between the simulation and biological evolution suggest that the model should only be thought of as conceptual exploration regarding the nature of evolutionary design. However, it makes precise the ways in which non-modularity can hinder mechanistic understanding, and helps to recognize the limits of the mechanistic model of explaining phenomena in the social and behavioral sciences.

## 5.6. Natural kinds and concept eliminativism

The last article takes a broader view on *the natural kinds model* of scientific concepts. In a slightly reflexive move, I examine the controversy surrounding the scientific concept of CONCEPT. Empirical research in psychology suggests that human conceptual abilities are sustained by several distinct cognitive mechanisms that do not act together in an orchestrated manner (see section 2.2.). That is, the cluster of properties typical of concepts is not sustained by one kind of mechanism, but rather by three different submechanisms: those corresponding to prototypes, exemplars, and theory-based concepts. According to Edouard Machery (2005; 2009; 2010a), the mechanistic theory of natural kinds suggests that the real natural kinds in this case are the ones corresponding to the submechanisms, and in order to avoid attribution errors, one should eliminate the notion of CONCEPT from scientific usage.

Although the empirical research underlying Machery's argument is relatively uncontroversial, his eliminativist conclusion has met a lot of resistance. The critics have argued that CONCEPT plays an indispensable role in psychological research. It captures a set of questions and generalizations having to do with the human capacity for conceptual thought in general. Abandoning the notion would therefore deemphasize this set and hinder scientific progress because there would be no notion to integrate results from research on subkinds of concepts (Couchman et al. 2010; Edwards 2010; Hampton 2010). Moreover,

defenders of the notion of concept in psychology have argued that despite being analyzable into further subkinds, CONCEPT qualifies as a mechanistic natural kind (Samuels & Ferreira 2010).

In my analysis, I draw two conclusions from this impasse. The view of conceptual change employed by Machery, the split-lump-eliminate scheme, (1) is not sufficient for concept eliminativism, and more fundamentally, (2) the natural kinds model of scientific concepts that it draws on is ambiguous and conflates several *epistemic roles* that a concept can play.

I argue that ultimately the different sides in the debate rely on different properties of natural kinds. According to my analysis, Machery emphasizes features of natural kind concepts that make them useful tools for referring to *open explananda*, whereas anti-eliminativists content themselves with a definition of natural kinds that allows them to function as mechanistically grounded *explanantia*. Both sides draw on the tradition of natural kinds: as has now been mentioned several times, the notion of kind has typically been used to separate inductively unexhausted concepts from stipulations, but an equally important epistemic role for natural kind terms has been to justify inductive inference regarding kind properties. I argue that these two epistemic roles are sustained by separate properties of kind terms, and that clearly distinguishing between these sets of properties brings forth a distinction between two types of scientifically pertinent concepts that I call *investigative kinds* and *instrument kinds*. Furthermore, I propose that in addition to these types of concepts, also non-mechanistic *framework kinds* play an important role in scientific research.

Treating a concept as an investigative kind means that in addition to justifying inductive inference, members of the kind are assumed to share yet unknown similarities, and thus we can learn more about them by empirically investigating the properties of their instances. For this reason, investigative kind concepts are good vehicles for representing targets of ongoing empirical research, and often stand for explananda in scientific theories.

On the other hand, instrument kinds can be internally heterogeneous in a way that often undermines their role as targets of further empirical study. There is no reason to assume that their members share properties apart from the ones governed by the homeostatic mechanism of the kind. In a sense, the unity of the kind is exhausted by knowing the finite cluster of kind-properties governed by the mechanism. Scientific concepts such as ENZYME, VITAMIN, and perhaps CONCEPT are based on a clear causally specified role, but are known to be structurally heterogeneous (Bechtel 1984; Couch 2009; Richardson 2008). However, as explanantia and as storages of scientific knowledge about phenomena, they are reliable: the abstractly defined causal mechanism of the kind supports a cluster of kind-typical properties, and therefore it is justified to assume that correctly identified members of the kind share this property cluster. I argue that such functionally identified but causally supported concepts are common in science and often face no fear of elimination.

Furthermore, I suggest that not all scientific concepts must be grounded in well-defined mechanisms. As pointed out already by the late logical empiricists, many prominent scientific kinds are cluster-notions (Putnam 1975a, 379). They are not defined by their role in a single law or theory, but reside at the intersection of several theories, and have slightly different meanings in different research programs. I suggest that despite not being anchored in any specific causal mechanisms, framework kinds like ENERGY, GENE, RATIONALITY or INFORMATION often play an important epistemic role. As suggested by Susan Leigh Star in her work on boundary objects, in science we need concepts simultaneously inhabiting several social worlds. They must be malleable enough to adapt to the informational requirements of different disciplines, but still maintain the identity of the target across different sites (Star & Griesemer 1989).

This distinction between different types of scientific concepts offers a plausible solution to the impasse between Machery and anti-eliminativists by suggesting that Machery's argument only amounts to

showing that CONCEPT is not an investigative kind. More generally, my approach draws attention to a hitherto overlooked form of conceptual change in science, *change in the inferential status* of the concept. In addition to the operations described by the SLE scheme, conceptual change consists also in often subtle changes in the inferential potential of concepts. The labels 'investigative,' 'instrumental,' and 'framework kind' correspond to such inferential statuses, and keeping track of how scientific concepts move from one concept-type to another is one way of representing such conceptual change.

## 6. Discussion

In this concluding section, I sketch the outlook on conceptual change in science suggested by the results obtained in the articles. Therefore, instead of aiming for a general summary of the sundry results, I focus on three features of the mechanistic theory that I find central for conceptual change: (1) the naturalness of mechanistic kinds forms a continuum; (2) natural kinds can be identified based on functional descriptions and thus have multiply realized components; and, (3) the mechanistic theory is compatible with classificatory pluralism. I propose that – despite their counter-intuitiveness – these properties together make the mechanistic theory a promising platform for theorizing about conceptual change in the social and behavioral sciences. In section 6.4., I draw the threads together and outline the picture of interdisciplinary conceptual integration suggested by my approach.

### 6.1. Kinds and mechanistic extrapolation

A central aim of this dissertation has been to argue that the mechanistic theory of natural kinds offers a plausible foundation for a naturalistic theory of concepts in the social and behavioral sciences. It captures a common practice, in which category-based inferences are justified by referring to knowledge of the causal structures underlying the phenomenon in question. Also, compared to the traditional empiricist models of induction in philosophy, the mechanistic theory provides a

richer picture of inductive inference, because it can be elaborated by connecting it to recent advancements in the philosophical thinking on scientific explanation: Rather than putting questions of induction in terms of the nebulous notions like projectibility and genuine lawlikeness, the mechanistic theory makes it possible to frame such inference as *mechanistic extrapolation* based on knowledge of the sustaining causal structures and environmental conditions (Guala 2005; Steel 2008).

However, while the mechanistic theory is intended to offer a common picture of concept formation across the special sciences, it need not lead to disregarding important dissimilarities between domains. There clearly are differences between the properties of biological and social mechanisms, and respectively, the natural kinds relying on such mechanisms also have different properties. Therefore, the mechanistic picture is no fixed procrustean bed that stretches concepts used in the social and behavioral sciences to fit a rigid model of classification originating in the natural sciences. Instead, my approach suggests that *naturalness of kinds comes in degrees*. Classic examples of natural kinds are characterized by a cluster of properties sustained by a stable and robust mechanism. For example, the typical properties of noble metals (e.g., reflectivity, density, malleability) are constituted by the lattice structure of the atoms. The structure is not sensitive to disturbances, and produces the same observable properties under a wide range of background conditions. By contrast, mechanisms underlying social phenomena are often fragile. For instance, the observable features of a social group do not generally depend only on the intrinsic properties of the group members themselves, but instead they result from interactions with non-members, and are therefore sustained by context-sensitive and extrinsic mechanisms. Moreover, as the looping effect of human kinds suggests, the functioning of the mechanisms of such interactive kinds can sometimes be altered simply by revealing their existence (Mannheim 1952). That is, when people become aware of classifications pertaining to them, their behavior changes.

Consequently, inductive inference regarding human kinds is more complicated and sensitive to disturbances than in the more traditional cases. However, as suggested in [II], the epistemic advantages of mechanistic extrapolation apply also in these situations involving agency and interpretation: Despite the interaction effects characteristic of human kinds, it is desirable that explanation and prediction of phenomena build on causal knowledge of the processes underlying such phenomena.

## 6.2. Classificatory strategies for complexity

An important aspect of a naturalistic theory of scientific concepts and conceptual change is that it should be compatible with what is known about the users of such representations. Therefore, the theory must somehow account for the gap between the limited cognitive capacities of those who use the representations and the often causally very complex nature of the phenomena studied (Mitchell 2003). I suggest that two features of my mechanistic theory, (1) its compatibility with abstract mechanisms and (2) classificatory pluralism, offer a response to this challenge.

As is pointed out in [VI], viable scientific concepts often refer to mechanistically heterogeneous structures. Kinds like EYE, ENZYME, or CONCEPT are individuated by the functional properties that they have as parts of causal systems, and on more fine-grained levels of description, the instances of these kinds can differ from each other in many ways. Sometimes the existence of such multiply realized kinds has been used as an argument against the need for a mechanistic grounding for special science kinds (Kincaid 1997; Weiskopf 2011). However, I hold that the mechanistic theory of natural kinds can account for the sensible anti-reductionist intuition behind the multiple realizability arguments while still holding on to the idea that the ability of kind-concepts to license inductive inferences is based on them being anchored in causal structures.

The crucial conceptual resource for this argumentative move is the non-fundamentalist understanding of mechanisms stemming from the contrastive-counterfactual theory of explanation. It makes it possible to understand functionally individuated *instrument kinds* as relying on abstractly characterized causal mechanisms [VI]. Instrument kinds can be regarded as being multiply realized in the sense that while instrument-kind concepts support invariant generalizations and interventions regarding the properties governed by the homeostatic mechanism, they remain agnostic about several implementation-level differences between the instances of the kind.

Now, I suggest that the motivation for instrument-kind classifications is the management of the cognitive economy. As has been argued by several authors, scientific representations are inevitably partial, and each representation can only capture some aspects of the complex targets of study (Cartwright 1999; Mitchell 2003, 183; Mäki 1994; Mäki 2011; Wimsatt 2007). In the same spirit, I propose that instrument-kind concepts could be seen as a result of a classificatory strategy that Levins (1966) called *sufficient parameters*: higher-level classifications are often created by intentionally leaving out parameter values not relevant for the epistemic project at hand. Instrument kinds thus result from the epistemic strategy of abstraction, and the kind of multiple realizability manifested by the resulting kinds does not create an ontological puzzle, unlike is often assumed in the literature on the topic (cf. Shapiro 2000).[14]

Another perhaps counter-intuitive property of the mechanistic theory of natural kinds uncovered in the articles is the *classificatory pluralism*

[14] This leaves open the possibility of there being other kinds of multiple realization, in which the kind-properties would result from widely heterogeneous underlying structures. However, as Bechtel and Mundale (1999) have suggested, such radical multiple realizability is probably more a philosopher's fiction than an empirically live possibility.

inherent in the theory. As argued in [IV], the mechanistic theory suggests that different scientific fields sharing a target of study can often demarcate the mechanisms underlying the target in different ways, and consequently, end up with different ways of classifying the domain.[15]

Instead of taking this as an argument against my position, I suggest that since pluralism is a fact of the matter in the sciences (Mitchell 2003), it is an advantage of my theory of conceptual change that it can account for the fact. Often the targets studied in the social and behavioral sciences are products of evolutionary design, and many of their properties are sustained by intricate causal interactions with the environment. In consequence, it becomes less obvious where the boundaries of such systems should be drawn, and such targets often refuse to be captured by a single classificatory scheme. I suggest that this complexity of targets is often reflected in scientific research practices [IV]. For example, in memory research, the laboratory and ecological paradigms delineate their systems of interest in different ways, and approach them with variables formulated at different levels of abstraction. This leads to the different approaches having what I call *different profiles of explanatory power*, i.e. they can answer different clusters of explanatory questions regarding the target phenomena.

I propose that like multiply realized kinds, this natural kinds pluralism should not be seen as a ontological conundrum, but instead as an epistemic strategy for dealing with complexity. By employing parallel classification schemes that can meet different inferential and practical demands, scientific disciplines studying causally complex phenomena bring about a division of cognitive labor: In order to produce finite and understandable theories of their targets, different scientific fields

[15] This pluralist tendency can be found also in Boyd's own theory of kinds (Boyd 1999) and in the recent formulations of his position, he has made this aspect of the theory more explicit than before (2010). Boyd's writings, however, offer little guidance for how to overcome the kind of relativism implied by indexing kind-terms to disciplinary matrices.

aim to capture different parts of the complex causal web sustaining the targets they study. By so doing, they can focus on those properties that are central to the epistemic aims of their particular discipline.

## 6.3. Beyond the natural kinds model

At this point, one could challenge my view by pointing out that I have watered down the idea of natural kind to the point where the resulting classifications are no longer *natural* in any meaningful sense. That is, because my theory does not clearly differentiate between universal and spatiotemporally restricted kinds, micro-level kinds and macro-kinds, naturally occurring and artificial kinds, or even between biological and social kinds, it seems to do away with most of the common contrasts intuitively associated with the naturalness of a kind.

While this objection is reasonable, I do not see it primarily as a challenge to my account of scientific concepts, but instead as a general argument against the use of the notion of natural kind. As noted by Hacking (2006) and Griffiths (2004), the concept is a loaded one. Because of the multitude of distinct contrasts to naturalness of a kind, the use of the notion can often spark unnecessary proprietary debates about its correct meaning, and create more confusion than clarity. However, problems with the concept of natural kind do not undermine the usefulness of the mechanistic theory as an account of conceptual change: None of the senses of naturalness mentioned in the paragraph above are necessary for guaranteeing the particular property of natural kinds that has motivated the epistemology-oriented tradition of kinds since Mill (2002) – the epistemic reliability of scientific concepts. As I argue on several occasions in the following articles, it is precisely the mechanistic grounding of natural kinds that separates them from arbitrary or conventional classifications that fail to support reliable inferential practices.

One way to avoid unnecessary confusions would be to regiment the use of the notion of natural kind so that only categories with intrinsic essences and crisp perspective-independent boundaries would

be called natural ones. Such kinds can mostly be found in physics and chemistry, and usually they tend to fall on the intuitively correct side of the contrasts listed above (cf. Ellis 2001). However, this does not in any way reduce the need for a theory of the foundations of concept formation in the social and behavioral sciences. On the contrary, I suspect that when faced with such causally messy domains, our everyday intuitions of where the phenomena are, and how the units of analysis should be chosen, tend to become increasingly unreliable [IV]. In such cases, having a systematic theory of classification and concept formation seems crucially important.

Therefore, I regard it as an important virtue of the mechanistic theory that it can provide a realist account of classification even for causally complex target domains. While the kind-formation strategies of abstraction and pluralism described above suggest that the epistemic aims of a discipline have a role in classificatory decisions, this makes mechanistic kinds conventional only in a weak sense. Despite the differences in the ways in which different disciplines describe their mechanistic variables and how they draw mechanism boundaries, the theory portrays the different classificatory perspectives as all latching on to the same objectively existing structures in the world. Ultimately, whether philosophers would call the resulting classifications natural kinds is not an issue of great importance to scientific practice.

In the following concluding section of this introductory essay, I make some brief remarks on how the mechanistic theory could be seen to form the core of a framework for thinking about conceptual change and interfield integration. I suggest that making explicit the mechanistic commitments of investigative kinds and instrument kinds as well as the non-mechanistic ones of framework kinds offers a way to see the different conceptualizations of targets in different scientific fields as setting inferential constraints on one another. Metaphorically, these constraints could be seen as the threads that weave the separate classificatory perspectives into parts of a more unified picture of the stud-

ied phenomena. The following discussion thus complements the aforementioned strategies of abstraction and pluralism with a third one required by interdisciplinary knowledge production, that of *conceptual integration*.

## 6.4. Conclusion. Weaving the mechanistic fabric

In 20th century philosophy, the most well-known way of thinking about integration between the sciences was epitomized by Oppenheim and Putnam's (1958) hierarchical view, according to which the unity of science was to be reached by expressing the vocabulary and the laws of the special sciences in the language of physics by means of a series of successive micro-reductions. However, since the 1960s, much of the discussion around the organization of the sciences has challenged this idea. Arguments based on multiple realizability, ontological disunity of the sciences, and methodological discontinuities between scientific disciplines have all aimed to show the insufficiency of the reductionist model (Fodor 1974; Dupré 1995; Cartwright 1999; Taylor 1971; Geertz 1973).[16]

In the first section of this introductory essay, I pointed out that examination of phenomena residing at interdisciplinary boundaries in the social and behavioral sciences suggests that neither the reductionist-unity picture nor the views defending strong autonomy of the special sciences can offer a satisfactory account of interdisciplinary epistemic collaboration. To return to an example discussed in section 1 and many of the articles, consider again Asperger's syndrome. It seems highly implausible that, in the near future, the explanatory resources of the "lowest-level" disciplines studying the human mind (i.e., cellular and molecular neuroscience) would be sufficient for a satisfactory explanation of the phenomenon – not to mention the design and implementation of effective interventions. In general, phe-

[16] For comprehensive discussions of the developments of the idea of the unity of science, see Cat (2010) and Bechtel & Hamilton (2007).

nomena in the social and behavioral sciences are often embedded in social and cultural structures in a way that strongly suggests that also higher-level factors must play a role in their explanation.

On the other hand, also positions advocating strong disconnections between the different fields in the social and behavioral sciences appear untenable. Given a sensible materialist picture of the human mind and sociality, it should be quite uncontroversial that ultimately the world of meanings and interpretation is supported by the mind-brain, and in consequence, more precise knowledge regarding the functioning of this system can be helpful also for explaining intentional and social behavior. Therefore, while the non-reductionist motivation underlying the disunity arguments is plausible, adherents of the non-reductionist views often go too far when they promise to insulate the conceptual systems of different disciplines from each other.

All in all, the unity-disunity dichotomy does not seem like a fruitful framing for questions of epistemic integration between the sciences, and strong claims of disciplinary autonomy in the social and psychological sciences have often been motivated not so much by theoretical or ontological reasons than by sociological ones having to do with the disciplinary identities of the social sciences (Wallerstein 1996, ch. 1). My approach to interfield coordination builds on a more egalitarian view of the organization of the sciences first envisioned by Otto Neurath (1937; 1946). Unlike the more well-known reductionist vision adopted by many of the other logical empiricists, Neurath conceived of the unity of science not as an ideal endpoint of inquiry, but as an ongoing process of iterative systematization of terminology, symbolism, and theoretical tools between different scientific fields. Angela Potochnik (2011) suggests that the work of authors such as Darden and Maull (1977), Bechtel (1984), Mitchell (2003), and Craver (2007) can be seen as advancing this *coordinate unity* view of the organization of the sciences.

For my purposes, especially Darden and Craver's work on interfield theories and multi-level mechanisms seems like a promising start-

ing point for thinking about conceptual integration. Craver's (2007, ch. 7) recent picture of the mosaic unity of the neurosciences represents probably the most developed version of this approach. Using the history of memory research in neuroscience as his case study, Craver proposes that by examining the relationships between mechanisms at several levels, it is possible to construct a multi-level mechanistic description of the target system. The gist of Craver's model is to understand integration not as reduction or derivation, but as *accumulation of mutual constraints* between mechanism descriptions. Roughly, lower-level accounts provide knowledge of the entities and activities occurring in the higher-level mechanisms, and higher-level mechanisms are often necessary for identifying the roles of these entities in the more general picture of the functioning of the system. As the number of such mutual constraints between levels increases, it becomes possible to fill in the details in the initial mechanism sketch, and the research converges towards a more accurate description of the actual multi-level mechanism sustaining the target phenomenon.

I propose that this approach to bridging mechanism descriptions through the identification of constraints between research perspectives is useful also for thinking about integration between distinct mechanism-based concepts. As suggested above in section 6.2., the mechanisms underlying parallel conceptualizations of shared targets often overlap and reside at different levels of abstraction. Therefore, the general task in conceptual integration is to work out the relationships between the various mechanism descriptions corresponding to concepts, and, if needed, modify the classification schemes so that conflicts between mechanistic commitments are resolved.

I think that both the bottom-up and top-down constraints suggested by the mosaic unity picture, as well as interfield relationships introduced by Darden and Maull (1977), provide useful heuristics for weaving the parallel mechanisms into parts of a more coherent whole: (i) identities; (ii) part-whole relationships; (iii) cause-effect relationships; and, (iv) function-structure relationships between mechanisms

or their components all suggest different ways in which mechanistic structures underlying the classificatory schemes of neighboring fields can be related. Ultimately the epistemic payoff from discovering such constraints is to make new information available for research in a scientific discipline, and thereby facilitate the creation of more accurate conceptualizations of targets. More precisely, explicating the mechanistic commitments of concept use should lead to the identification of at least two kinds of *inferential relationships* between concepts in neighboring domains, those of conflict and complementarity.

First, explication of the mechanistic commitments of concept use can lead to perception of genuine conflict*s* between classification schemes, suggesting that at least one of them must be revised. For example, in the psychological research on concepts, mechanistic conflicts between the neo-empiricist approach and the more traditional prototype, exemplar, and theory-based theories seem likely. The different theories are phrased at approximately the same level of abstraction, and aim to explain roughly the same explananda. However, they make substantially different presuppositions concerning the processes and representations sustaining our capacity of conceptual thought. A promising strategy for resolving the conflict would be to look to fields like cognitive neuroscience and neuropsychology for additional mechanistic information about the processes underlying the competing theories, and based on that information determine which theory must give way.[17]

However, the non-fundamentalist approach to mechanisms introduced in section 4.3. suggests that *prima facie* competing conceptualizations can also often turn out to be compatible, and often complementary. The epistemic strategies of abstraction and classificatory pluralism allow neighboring fields to end up with mutually enriching classificatory strategies, where the dissimilarities between the mechanisms

[17] See, for example, Machery (2010b) for references to possibly relevant research on the neuroscience of concepts.

employed reflect the division of cognitive labor. As argued in [IV], different mechanism descriptions can be used to answer distinct contrastive explananda, and illuminate different aspects of the target. For example, such a solution has recently been proposed to perennial disputes regarding the causal efficacy of personal-level psychological states. As argued by proponents of the contrastive-counterfactual theory, neuroscientific and personal-level explanations of behavior typically capture different contrastive-counterfactual dependencies and therefore need not be incompatible with each other (Raatikainen 2010; Woodward 2008; Ylikoski 2001).

In addition to these mechanism-based forms of conceptual integration, the approach to scientific concepts developed in this dissertation also goes beyond the mosaic unity model. I have claimed that causally-based concepts do not exhaust the variety of concepts used in science. Instead, I proposed that at least three types of scientific concepts are employed in research: investigative kinds that refer to open explananda with yet unexhausted inductive potential, instrument kinds that act as reliable explanantia, and framework-kind concepts, which are not anchored in any particular causal mechanism.

As I argue in [VI], rather than invalidating the mechanistic theory, acknowledging the existence of these multiple ways of functioning of concepts merely places the theory within a more general picture of the conceptual change. Anchoring scientific concepts in well-specified causal structures roughly in the way suggested by the SLE scheme helps to make them inferentially reliable, and as suggested in the paragraphs above, examining the interrelations between mechanistically grounded concepts provides a means for obtaining more encompassing knowledge of the complex targets of study. However, thanks to their particular way of functioning, framework concepts appear to serve a complementary role, where they act as the communicative glue in the mosaic of interdisciplinary research, bringing together the different perspectives.

As a tentative suggestion, my approach thus recommends that in addition to tracing the mechanistic commitments of concept use, a theory of conceptual change in science should pay attention to the various epistemic roles that concepts can have in research. By keeping track of the different types of inferential commitments made by using concepts, it is possible to facilitate interfield communication, guard against error, and weave contributions from distinct fields into parts of a coherent and comprehensive picture of the target of research. While this picture of interfield integration is inevitably a mere sketch, and much remains to be done in future work on the topic, I believe that the approach developed in this dissertation strikes a plausible balance between the unity and disunity of science. My theory of conceptual change does not envision the reduction of the patchwork of theories and concepts to a single representational framework, but instead it suggests that integrative efforts involve groping in the dark for reliable inferential connections between concepts employed in different fields. However, it does retain integration as a theoretical ideal for scientific research in a world that is at the same time disordered, dappled, but one.

## References

Arguello, A., & Gogos, J. (2012). Genetic and cognitive windows into circuit mechanisms of psychiatric disease. *Trends in Neurosciences,* 35, 3–13.

Ayers, M. (1981). Locke versus Aristotle on natural kinds. *Journal of Philosophy,* 78, 247–272.

Barker, M. (2010). From cognition's location to the epistemology of its nature. *Cognitive Systems Research,* 11, 357–366.

Baron-Cohen, S. (2000). Theory of mind and autism: A review. In L. M. Glidden (ed.), *International Review of Research in Mental Retardation.* Vol. 23 (pp. 169–184). Academic Press.

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences,* 22, 577–660.

Bechtel, W. (1984). Reconceptualizations and interfield connections: The discovery of the link between vitamins and coenzymes. *Philosophy of Science,* 51, 265–292.

— (2008). Mechanisms in cognitive psychology: What are the operations? *Philosophy of Science,* 75, 983–994.

— (2009). Looking down, around, and up: Mechanistic explanation in psychology. *Philosophical Psychology,* 22, 543–564.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences,* 36, 421–441.

Bechtel, W., & Hamilton, A. (2007). Reduction, integration, and the unity of science: Natural, behavioral, and social sciences and the humanities. In T. Kuipers (ed.), *Philosophy of Science: Focal Issues* (Volume 1 of the Handbook of the Philosophy of Science). Elsevier.

Bechtel, W., & Mundale, J. (1999). Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science,* 66, 175–207.

Bechtel, W., & Richardson, R. (2010). *Discovering Complexity: Decomposition and localization as strategies in scientific research*. Cambridge, MA: MIT Press.

Bennett, M., Dennett, D., Hacker, P., & Searle, J. (2007). *Neuroscience and Philosophy: Brain, Mind, and Language*. New York: Columbia University Press.

Bennett, M. R., & Hacker, P. M. S. (2003). *Philosophical Foundations of Neuroscience*. Oxford: Blackwell.

— (2008). *History of Cognitive Neuroscience*. Oxford: Wiley-Blackwell.

Berlin, B. (1992). *Ethnobiological Classification: Principles of categorization of plants and animals in traditional societies*. Princeton: Princeton University Press.

Bermudez, J. L. (2004). *Philosophy of Psychology: A contemporary introduction*. London: Routledge.

Bird, A., & Tobin, E. (2008). Natural kinds. In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Summer 2010 Edition)*, URL = <http://plato.stanford.edu/archives/sum2010/entries/natural-kinds/>

Bishop, M. A. (2002). The theory theory thrice over: The child as scientist, superscientist, or social institution? *Studies in History and Philosophy of Science,* 33, 121–36.

Block, N. (1998). Conceptual role semantics. In E. Craig (ed.) *Routledge Encyclopedia of Philosophy*. Routledge.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *Philosophical Review,* 97, 303–352.

Bourdieu, P., & Wacquant, L. (1992). *An Invitation to Reflexive Sociology*. Chicago: University Of Chicago Press.

Boyd, R. (1980). Scientific realism and naturalistic epistemology. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1980*, 613–662.

— (1983). On the current status of the issue of scientific realism. *Erkenntnis,* 19, 45–90.

— (1991). Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies*, 61, 127–148.

— (1999). Kinds as the 'Workmanship of Men.' In J. Nida-Rümelin (Ed.), *Rationalität, Realismus, Revision* (pp. 52–89). Berlin: Walter de Gruyter.

— (2010). Realism, natural kinds, and philosophical methods. In H. Beebee, & N. Sabbarton-Leary (eds.), *The Semantics and Metaphysics of Natural Kinds*. Routledge.

Brandom, R. (1994). *Making it Explicit: Reasoning, representing, and discursive commitment*. Cambridge, MA: Harvard University Press.

Brigandt, I. (2010). The epistemic goal of a concept: accounting for the rationality of semantic change and variation. *Synthese,* 177, 19–40.

Brock, J. (2012). Alternative Bayesian accounts of autistic perception: comment on Pellicano and Burr. *Trends in Cognitive Sciences,* 16, 573–574.

Cacioppo, J., & Decety, J. (2011). Social neuroscience: challenges and opportunities in the study of complex behavior. *Annals of the New York Academy of Sciences,* 1224, 162–173.

Carey, S. (2009). *The Origin of Concepts*. Oxford: Oxford University Press.

Carnap, R. (1932). The elimination of metaphysics through logical analysis of language. *Erkenntnis*, 60–81.
— (1950). empiricism, semantics, and ontology. *Revue Internationale de Philosophie,* 4, 20–40.
Cartwright, N. (1999). *The Dappled World: A Study of the boundaries of science.* Cambridge, MA: Cambridge University Press.
Cat, J. (2010). The unity of science. In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*.
Churchland, P. (1984). *Matter and Consciousness: A contemporary introduction to the philosophy of mind.* Cambridge, MA: MIT.
— (2005). Cleansing Science. *Inquiry,* 48, 464–477.
Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again*. Cambridge, MA: MIT Press.
— (2007). Curing cognitive hiccups: A defense of the extended mind. *Journal of Philosophy,* 104, 163–192.
— (2008*). Supersizing the Mind: Embodiment, action, and cognitive extension.* Oxford: Oxford University Press.
Couchman, J., Boomer, J., Coutinho, M. V. C., & Smith, J. D. (2010). Carving nature at its joints using a knife called concepts. *Behavioral and Brain Sciences,* 33, 207–208.
Couch, M. (2009). Multiple realization in comparative perspective. *Biology and Philosophy,* 24, 505–519.
Craver, C. (2004). Dissociable realization and kind splitting. *Philosophy of Science,* 71, 960–971.
— (2007). *Explaining the Brain: Mechanisms and the mosaic unity of neuroscience*. Oxford: Clarendon.
— (2009). Mechanisms and natural kinds. *Philosophical Psychology,* 22, 575–594.
Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin,* 52, 281–302.
Cummins, R. (1983). *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press.
Darden, L. & Maull, N. (1977). Interfield theories. *Philosophy of Science,* 44, 43–64.

Dennett, D. (2007). Philosophy as naive anthropology: Comment on Bennett and Hacker. In Bennett et al., *Neuroscience and Philosophy: Brain, Mind, and Language*. Columbia University Press.

Downes, S. M. (1999). Can scientific development and children's cognitive development be the same process? *Philosophy of Science,* 66, 565–578.

Dretske, F. (1981). *Knowledge and the Flow of Information*. Oxford: Blackwell.

Duchaine, B., Cosmides, L., & Tooby, J. (2001). Evolutionary psychology and the brain. *Current Opinion in Neurobiology,* 11, 225– 230.

Dunbar, R. (1998). The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews,* 6, 178–190.

Dupré, J. (1995). *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*. Cambridge, MA: Harvard University Press.

Edwards, K. (2010). Unity amidst heterogeneity in theories of concepts. *The Behavioral and Brain sciences,* 33, 210–211.

Eliasmith, C. (2005). Neurosemantics and categories. In H. Cohen, & C. Lefebvre (eds.) *Handbook of Categorization in Cognitive Science*. Amsterdam: Elsevier Press.

Elster, J. (2007). *Explaining Social Behavior: More nuts and bolts for the social sciences*. Cambridge: Cambridge University Press.

Evans, J., & Frankish, K. (2009). *In Two Minds: Dual processes and beyond.* Oxford: Oxford University Press.

Eyal, G. (2010). *The Autism Matrix*. 1st ed. Cambridge, MA: Polity.

Faucher, L. et al. (2002). The baby in the lab-coat: why child development is not an adequate model for understanding the development of science. In P. Carruthers, S. Stich, & M. Siegal, (eds.), *The Cognitive Basis of Science* (pp. 335–362). Cambridge, MA: Cambridge University Press.

Fodor, J. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese* 28, 97–115.

— (1983). *The Modularity of Mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.

— (1998). *Concepts: Where cognitive science went wrong*. Oxford: Clarendon Press.

— (2010). *LOT 2: The language of thought revisited*. Oxford: Oxford University Press, USA.
Frege, G. [1892] (2010). On sense and reference. In D. Byrne, & M. Kolbel (eds.) *Arguing About Language*. Routledge.
Garfinkel, A. (1981). *Forms of Explanation: Rethinking the questions in social theory*. New Haven: Yale University Press.
Geertz, C. (1973). Thick description: Toward an interpretive theory of culture. In Geertz, *The Interpretation of Cultures* (pp. 3–33). New York: Basic Books.
Gelman, S. (2003). *The Essential Child: Origins of essentialism in everyday thought*. Oxford: Oxford University Press.
Giddens, A. (1984). *The Constitution of Society*. Cambridge: Polity Press.
Gigerenzer, G., Hertwig, R., & Pachur, T. (2011). *Heuristics: The foundations of adaptive behavior*. New York: Oxford University Press.
Glennan, S. (2002). Rethinking mechanistic explanation. *Proceedings of the Philosophy of Science Association,* 2002, 342–353.
— (2005). Modeling mechanisms. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences,* 36, 443–464.
Glimcher, P., & Rustichini, A. (2004). Neuroeconomics: The consilience of brain and decision. *Science*, 306, 447–452.
Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA: Addison Wesley.
Goodman, N. [1955] (1983). *Fact, Fiction, and Forecast*. 4th. Cambridge, MA: Harvard University Press.
Gopnik, A. & Meltzoff, A. (1997). *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.
Grene, M., & Depew, D. (2004). *The Philosophy of Biology: An episodic history*. Cambridge: Cambridge University Press.
Griffiths, P. (1998). *What Emotions Really Are: The Problem of Psychological Categories*. Chicago: University of Chicago Press.
— (2004). Emotions as natural and normative kinds. *Philosophy of Science,* 71, 901–911.
Griffiths, P. & Stotz, K. (2008). Experimental philosophy of science. *Philosophy Compass,* 3, 507–521.

Guala, F. (2005). *The Methodology of Experimental Economics*. Cambridge, MA: Cambridge University Press.
Hacking, I. (1991). A tradition of natural kinds. *Philosophical Studies,* 61, 109–26.
— (1993). Working in a new world: The taxonomic solution. In Horwich, P. (ed.), *World Changes*. Cambridge, MA: MIT Press.
— (1995a). *Rewriting the soul: Multiple personality and the sciences of memory*. Princeton, N.J.: Princeton University Press.
— (1995b). The looping effects of human kinds. In D. Sperber, D. Premack, & J. Premack (eds.), *Causal Cognition. A multidisciplinary debate* (pp. 351–94). Oxford: Clarendon Press.
— (1998). *Mad travelers: Reflections on the reality of transient mental illnesses.* Charlottesville: University Press of Virginia.
— (1999). *The Social Construction of What*? Cambridge, MA: Harvard University Press.
— (2002). *Historical Ontology*. Cambridge, MA: Harvard University Press.
— (2006). Des classifications naturelles. Cours B: Les choses, les gens et la raison. Lecture at College de France. Retrieved July 15, 2012. (http://www.college-de -france.fr/media/historique/UPL32428_classifications_naturelles.pdf).
Hampton, J. (2010). Concept talk cannot be avoided. *The Behavioral and brain sciences,* 33, 212–213.
Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology,* 36, 49–67.
Hempel, C. (1965). *Aspects of Scientific Explanation.* New York: Free Press.
Hitchcock, C. (1995). Salmon on explanatory relevance. *Philosophy of Science,* 62, 304–320.
Holland, J. (1975). *Adaptation in Natural and Artificial Systems: An introductory analysis with applications to biology, control, and artificial intelligence*. Ann Arbor: University of Michigan Press.
Hornsby, J. (2000). Personal and sub-personal: A defence of Dennett's early distinction. *Philosophical Explorations,* 3, 6–24.
Hull, C. (1920). Quantitative aspects of the evolution of concepts. *Psychological monographs,* XXVIII.

Hull, D. (1980). On human nature. *Environmental Ethics,* 2, 81–88.
Hutchins, E. (1995). *Cognition in the Wild*. Cambridge, MA: MIT Press.
— (2010). Cognitive ecology. *Topics in Cognitive Science,* 2, 705–715.
Jacob, F. (1977). Evolution and tinkering. *Science,* 196, 1161–1166.
Jylkkä, J. (2008). *Concepts and Reference. Defending a dual theory of natural kind concepts*. Doctoral dissertation. Turku: University of Turku.
Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.
Keil, F. (1992). *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: A Bradford Book.
Kendler, K., Zachar, P., & Craver, C. (2010). What kinds of things are psychiatric disorders? *Psychological medicine*, 41, 1–8.
Kim, J. (1992). Multiple realization and the metaphysics of reduction. *Philosophy and Phenomenological Research,* 52, 1–26.
Kincaid, H. (1997). *Individualism and the Unity of Science*. Rowman Littlefield Publishers.
Kornblith, H. (1993). *Inductive Inference and its Natural Ground: An essay in naturalistic epistemology*. Cambridge, MA: MIT Press.
Kripke, S. (1980). *Naming and Necessity*. Oxford: Blackwell.
Kuhn, T. (1996). *The Structure of Scientific Revolutions*. 3rd. Chicago: University of Chicago Press.
Kuorikoski, J. (2010). *Society by Numbers. Studies on Model-Based Explanations in the Social Sciences*. Doctoral dissertation. Helsinki: University of Helsinki.
Laland, K., & Brown, G. (2002). *Sense and Nonsense: Evolutionary Perspectives on Human Behaviour*. Oxford: Oxford University Press.
Laurence, S., & Margolis, E. (1999). Concepts and cognitive science. In E. Margolis, S. Laurence (eds.) *Concepts: Core Readings*. Cambridge, MA: MIT Press.
Levins, R. (1966). Strategy of model building in population biology. *American Scientist,* 54, 421–31.
Locke, J. [1690](1975*). An Essay Concerning Human Understanding*. London: Oxford University Press.
Lombrozo, T. (2009). Explanation and categorization: How "why?" informs "what?" *Cognition,* 110, 248–253.

Lycan, W. (1990). The continuity of levels in nature. In W. Lycan (ed.), *Mind and Cognition. An anthology* (pp. 77–97). Oxford: Wiley-Blackwell.

Machamer, P., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science,* 67, 1–25.

Machery, E. (2005). Concepts are not a natural kind. *Philosophy of Science,* 72, 444–467.

— (2009). *Doing Without Concepts*. New York: Oxford University Press.

— (2010a) Precis of Doing Without Concepts. *Behavioral and Brain Sciences,* 33, 195–244.

— (2010b). Reply to Barbara Malt and Jesse Prinz. *Mind and Language*, 25, 634–646.

Mäki, U. (1994). Isolation, idealization and truth in economics. *Poznán studies in the philosophy of the sciences and the humanities* 38, 147–168.

— (2011). Models and the locus of their truth. *Synthese,* 180, 47–63.

Mallon, R. (2003). Social construction, social roles and stability. In F. Schmitt (ed.), *Socializing Metaphysics: The Nature of Social Reality* (pp. 327–354). Lanham, MD: Rowman and Littlefield.

Mannheim, K. (1952). Das Problem einer Soziologie des Wissens. In Mannheim, *Essays on the Sociology of Knowledge*. London: Routledge Kegan Paul.

Marcus, G. (2008). *Kluge: The haphazard construction of the human mind.* London: Faber.

Margolis, E., & Laurence, S. (eds.) (1999). *Concepts: Core Readings.* Cambridge, MA: A Bradford Book.

Margolis, E., & Laurence, S. (2012). Concepts. In N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy.* URL = <http://plato.stanford.edu/archives/fall2012/entries/concepts

McDowell, J. (1996). *Mind and World*. Cambridge, MA: Harvard University Press.

Millikan, R. (1999). Historical kinds and the "special sciences." *Philosophical Studies,* 95, 45–65.

— (2000). *On Clear and Confused Ideas: An essay about substance concepts.* Cambridge: Cambridge University Press.

Mill, J. S. [1891] (2002). *A System of Logic: Ratiocinative and Inductive.* Honolulu, Hawaii: University Press of the Pacific.

Mitchell, M. (1998). *An Introduction to Genetic Algorithms.* Cambridge, MA: A Bradford Book.

— (2011). *Complexity: A Guided Tour*. Oxford: Oxford University Press.

Mitchell, S. (2003). *Biological Complexity and Integrative Pluralism*. Cambridge, MA: Cambridge University Press.

Montague, P. R. et al. (2012). Computational psychiatry. *Trends in Cognitive Sciences,* 16, 72–80.

Murphy, D. (2006). *Psychiatry in the Scientific Image*. Cambridge, MA: MIT Press.

Murphy, G. (2004). *The Big Book of Concepts*. Cambridge, MA: A Bradford Book.

Murphy, G., & Medin, D. (1985). The role of theories in conceptual coherence. *Psychological Review,* 92, 289–316.

Nagel, E. (1961). *The Structure of Science: Problems in the Logic of Scientific Explanation.* Harcourt, Brace & World.

Neurath, O. (1937). Unified science and its encyclopaedia. *Philosophy of Science,* 4, 265–277.

— (1946). The orchestration of the sciences by the encyclopedism of logical empiricism. *Philosophy and Phenomenological Research,* 6, 496–508.

Oppenheim, P. & Putnam, H. (1958). Unity of science as a working hypothesis. In H. Feigl et al. (eds.), *Minnesota Studies in the Philosophy of Science*, vol. 2, Minneapolis: Minnesota University Press.

Peacocke, C. (1992). *A Study of Concepts*. Cambridge, MA: MIT Press.

Piccinini, G. & Scott, S. (2006). Splitting concepts. *Philosophy of Science,* 73, 390–409.

Potochnik, A. (2011). A Neurathian conception of the unity of science. *Erkenntnis,* 74, 305–319.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind*? Behavioral and Brain Sciences* 1, 515–526.

Prinz, J. (2002). *Furnishing the Mind: Concepts and their perceptual basis.* Cambridge, MA: MIT Press.

Putnam, H. (1975a). The analytic and synthetic. In H. Putnam, *Mind, Language and Reality: Philosophical Papers. Vol. 2.*(pp. 33–69). Cambridge University Press.

— (1975b). The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science,* 7, 131–193.
Quine, W. V. O. (1969). Natural kinds. In W.V.O. Quine, *Ontological Relativity and Other Essays* (pp. 114–38). Columbia University Press.
Raatikainen, P. (2010). Causation, exclusion, and the special sciences. *Erkenntnis,* 73, 349–363.
Reydon, T. (2009). Do the life sciences need natural kinds? *Croatian Journal of Philosophy,* 9, 167–190.
Richardson, R. (2008). Autonomy and multiple realization. *Philosophy of Science,* 75, 526–536.
Rosch, E. (1978). Principles of categorization. In E. Rosch, & B. Lloyd (eds.), *Cognition and Categorization* (pp. 27–48). Hillsdale, NJ: Lawrence Erlbaum.
Rosch, E., & Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology,* 7, 573– 605.
Rosenberg, A. (2005). Lessons from biology for philosophy of the human sciences. *Philosophy of the Social Sciences,* 35, 3–19.
Roth, I., & Barson, C. (2010). *The Autism Spectrum in the 21st Century: Exploring psychology, biology and practice*. London: Jessica Kingsley.
Ruben, D-H. (1990). *Explaining Explanation*. London: Routledge.
Rupert, R. (2004). Challenges to the hypothesis of extended cognition. *Journal of Philosophy,* 101, 389–428.
Rusanen, A-M., & Pöyhönen, S. (2013). Concepts in change. *Science & Education*, 22, 1389–1404.
Russell, B. (1948). *Human Knowledge, its Scope and Limits*. London: Allen and Unwin.
Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton, N.J.: Princeton University Press.
— (1990). *Four Decades of Scientific Explanation*. Minneapolis: University of Minnesota Press.
— (1998). *Causality and Explanation*. Oxford: Oxford University Press.
Samuels, R. (2009). Delusions as a natural kind. In M. Broome, & L. Bortolotti (eds.), *Psychiatry as Cognitive Neuroscience: Philosophical perspectives* (pp. 49–82). Oxford: Oxford University Press.
Samuels, R., & Ferreira, M. (2010). Why don't concepts constitute a natural kind? *The Behavioral and brain sciences,* 33, 222–223.

Sellars, W. (1963). Philosophy and the scientific image of man. In W. Sellars, *Science, Perception, and Reality* (pp. 35–78). Humanities Press/Ridgeview.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin.

Shapiro, L. (2000). Multiple realizations, *Journal of Philosophy, 97,* 635–654.

Simon, H. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society,* 106, 467–482.

Singh, I., & Rose, N. (2009). Biomarkers in psychiatry. *Nature,* 460, 202–207.

Smelser, N., & Baltes, B. (eds.) (2001). *International Encyclopedia of the Social and Behavioral sciences*. Amsterdam: Elsevier.

Smith, E., & Medin, D. (1981). *Categories and Concepts*. Cambridge, MA: Harvard University Press.

Sober, E. (1980). Evolution, population thinking, and essentialism. *Philosophy of Science,* 47, 350–383.

Sperber, D. (1996). *Explaining Culture: A naturalistic approach*. Oxford: Blackwell.

Sprevak, M. (2010). Inference to the hypothesis of extended cognition. *Studies in History and Philosophy of Science Part A* 41, 353–362.

Stalker, D. (1994). *Grue!: The new riddle of induction*. Chicago: Open Court.

Star, S. L. & Griesemer, J. (1989). Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's museum of vertebrate zoology, 1907-39. *Social Studies of Science,* 19, 387–420.

Steel, D. (2008). *Across the Boundaries: Extrapolation in Biology and Social Science*. Oxford: Oxford University Press.

Sterelny, K. (2003). *Thought in a Hostile World: The evolution of human cognition*. Malden, MA: Blackwell.

Stich, S. (1996). *Deconstructing the Mind*. New York: Oxford University Press.

Taylor, C. (1971). Interpretation and the sciences of man. *The Review of Metaphysics,* 25, 3–51.

Uttal, W. (2003). *The New Phrenology: The Limits of Localizing Cognitive Processes in the Brain*. Cambridge, MA: A Bradford Book.

van Fraassen, B. (1980). *The Scientific Image*. Oxford: Oxford University Press.

Vosniadou, S. (2008). *International Handbook of Research on Conceptual Change*. Taylor & Francis.

Wallerstein, I. (1996). *Open the Social Sciences: Report of the Gulbenkian commission on the restructuring of the social sciences*. Stanford, CA: Stanford University Press.

Waskan, J. (2011). Mechanistic explanation at the limit. *Synthese,* 183, 389–408.

Weiskopf, D. (2011). Models and mechanisms in psychological explanation. *Synthese,* 183, 313–338.

Whewell, W. [1847] (1967). *The Philosophy of the Inductive Sciences*. 2nd. London: Cass.

Wilson, R. (1999). *Species: New interdisciplinary essays*. Cambridge, MA: MIT Press.

Wilson, R., Barker, M., & Brigandt, I. (2007). When traditional essentialism fails. *Philosophical Topics,* 35, 189–215.

Wimsatt, W. (2007). *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge, MA: Harvard University Press.

Wittgenstein, L. (1953). *Philosophical Investigations*. New York, Macmillan.

Woodward, J. (2002). What is a mechanism? A counterfactual account. *Proceedings of the Philosophy of Science Association* 2002, 366–377.

— (2003). *Making Things Happen: A theory of causal explanation*. New York: Oxford University Press.

— (2008). Cause and explanation in psychiatry: An interventionist perspective. In K. Kendler, & J. Parnas (eds.), *Philosophical Issues in Psychiatry: Explanation, Phenomenology, and Nosology* (pp. 132–183). Johns Hopkins University Press.

Wright, C., & Bechtel, W. (2007). Mechanisms and psychological explanation. In P. Thagard (ed.), *Philosophy of Psychology and Cognitive Science* (pp. 31–80). Elsevier.

Ylikoski, P. (2001*). Understanding Interests and Causal Explanation.* PhD thesis. University of Helsinki.

— (2007). The idea of contrastive explanandum. In J. Persson & P. Ylikoski (eds.), *Rethinking Explanation* (pp. 27–42). Dordrecht: Springer.

— (2011). Social mechanisms and explanatory relevance. In P. Demeulenaere (ed.), *Analytical Sociology and Social Mechanisms* (pp. 154–172). Cambridge: Cambridge University Press.

— (2012). Micro, macro, and mechanisms. In H. Kincaid (ed.), *The Oxford Handbook of Philosophy of Social Science* (pp. 21–45). Oxford: Oxford University Press.

Ylikoski, P., & Kuorikoski, J. (2010). Dissecting explanatory power. *Philosophical Studies,* 148, 201–219.

Zola-Morgan, S. (1995). Localization of brain function: The legacy of Franz Joseph Gall (1758-1828). *Annual Review of Neuroscience,* 18, 359–383.