Concepts Enacted: Confronting the Obstacles and Paradoxes Inherent in Pursuing a Scientific Understanding of the Building Blocks of Human Thought

Joel Parthemore

Submitted for examination in the degree of Doctor of Philosophy
University of Sussex, UK
September 2010

Declaration

I hereby declare that this thesis has not been and will not be submitted, in whole or in part, to this or any other University for the award of any other degree.

Signature: _____

Summary

University of Sussex
Joel Edward Parthemore

Concepts Enacted:

Confronting the Obstacles and Paradoxes Inherent in Pursuing a Scientific Understanding of the Building Blocks of Human Thought

This thesis confronts a fundamental shortcoming in cognitive science research: a failure to be explicit about the theory of concepts underlying cognitive science research and a resulting failure to justify that theory philosophically or otherwise. It demonstrates how most contemporary debates over theories of concepts divide over whether concepts are best understood as (mental) representations or as non-representational abilities. It concludes that there can be no single correct ontology, and that both perspectives are logically necessary. It details three critical distinctions that are frequently neglected: between concepts as we possess and employ them non-reflectively, and concepts as we reflect upon them; between the private (subjective) and public (inter-subjective) aspects of concepts; and between concepts as approached from a realist versus anti-realist perspective. Metaphysical starting points fundamentally shape conclusions.

The main contribution of this thesis is a pragmatic, meticulously detailed, and distinctive account of concepts in terms of their essential nature, core properties, and context of application. This is done within the framework of Peter Gärdenfors' conceptual spaces theory of concepts, which is offered as a bridging account, best able to tie existing theories together into one framework. A set of extensions to conceptual spaces theory, called the unified conceptual space theory, are offered as a means of pushing Gärdenfors' theory in a more algorithmically amenable and empirically testable direction. The unified conceptual space theory describes how all of an agent's many different conceptual spaces, as described by Gärdenfors, are mapped together into one unified space of spaces, and how an analogous process happens at the societal level.

The unified conceptual space theory is put to work offering a distinctive account of the co-emergence of concepts and experience out of a circularly causal process. Finally, an experimental application of the theory is presented, in the form of a simple computer program.

Acknowledgments

My biggest thanks go to my primary supervisor, Blay Whitby, for his unwavering support through difficult circumstances; his unwavering belief in my research and in my thesis; his detailed and patient reading of my thesis drafts; and his refusal, at any number of points, to allow me to give up. I'm glad to count you as both supervisor and friend. Huge thanks go as well to Kerith Harris, who as student adviser helped me through one of the most difficult periods I have known academically or personally. Beyond that, you were a sounding board, a voice of calm and reason, a source of advice and perspective, an intermediary, a welcoming shoulder.

Thanks to my secondary supervisor, Chris Thornton, who came on board late in the game, as I was coming close to the end of my degree; and who always pressed me, whenever we met, to say what my main theoretical contribution was meant to be.

Thanks to the third member of my research committee, Steve Torrance, whom I originally intended for my primary supervisor. You gave me my first introduction to the philosophical side of cognitive science back in my undergraduate course, in a class titled Computer Models of Mind.

Big thanks to Peter Gärdenfors for hosting my stay at Lund University while writing up my thesis, serving as an informal "second supervisor", and offering insightful feedback on many of my thesis chapters. Your humility and your humour both served as inspiration.

A big round of applause to my sister Judy for her assistance proofreading the page proofs for the *Pragmatics & Cognition* article that summarizes chapters Six and Seven, and for reading through the final thesis for spelling, punctuation, and referencing errors. Thanks as well to:

. . . Inman Harvey for providing me the key insights on the nature of symbols and representations in Chapter Two. I know I'm not the first doctoral student you helped set straight. Your qualified endorsement of my own position on representations means more to me than most of the other words of encouragement I've received. I find it amusing that the "big, bad anti-representationalist" Inman Harvey managed to make me sound more "anti-representationalist" than he!

. . . Björn Sjöden, Paulina Lindström, Åsa Harvard, Andreas Lind, Rasmus Bååth, and all the members of the cognitive science section and attendees of the cognitive science seminar in the department of philosophy at Lund University. Paulina, you always had written comments on my chapters, and I took most of them to heart. Björn, your friendship and our conversations helped me through a number of rough spots. More than anyone in Lund, you made me feel at home. *Tack så mycket!*

. . . Christian Balkenius, and all the faculty of the department of philosophy, for welcoming me.

. . . Cathrine Felix and Björn Petersson, for setting me straight on Davidson's distinction between actions and events; Carlo Proietti and Frank Zenker, for your help with my response to Penrose.

. . . Fritz-Anton Fritzson and the other members of the board of *Filosofiska Föreningen* ("Philosophical Society") for your support and friendship.

. . . Mathias Osvath for helping me understand better the conceptual and generally amazing cognitive abilities of the great apes (and corvids!).

. . . Jordan Zlatev and Göran Sonesson for your comments on my paper on the evolution of concepts (which forms part of Chapter Four), and for offering me the postdoctoral position in the Center for Cognitive Semiotics. Thanks as well to Johan Blomberg for comments on that paper.

. . . Araneya Olausson and Torleif Persson for comments that were often more useful than those I received from my fellow doctoral students or professors. Getting quality feedback on my writing – getting anyone to read my chapters in the first place – has been, I have discovered, one of the most difficult tasks of being a doctoral student. I am not sure though, Torleif, if I got rid of all the double negatives that annoyed you so much!

. . . Ron Chrisley, for reminding me that clear is never clear enough, and for some particularly useful comments on Chapter Three.

. . . Mike Beaton, Stan Rosenthal, Leslie Marsh, Tom Froese, and all the members of the Philosophy of AI in Cognitive Science (PAICS) research group at the University of Sussex. If I can survive your questioning, I can handle any audience!

. . . Tony Morse, my co-author on the *Pragmatics & Cognition* article. Your invitation to spend some time in Skövde writing a joint paper could not have come at a better moment; without that opportunity, I might well have abandoned my thesis. Meanwhile, if it hadn't been for you, I never would have ended up in Sweden!

. . . Alan Batie and Gene Lunderman for helping bail me out financially more than once when student loans or reimbursement monies were delayed.

. . . Peter Crowther for remembering Charley and asking me whether "Charley lives".

. . . Peter Timmerman for a particularly memorable late-night cycle ride through Groningen.

. . . My mother, for her support. Provided I pass, I hope you can make it to my graduation ceremony!

Thanks as well for helping me in one way or another through the last five years to Jakob Ahlin, Caroline Ailanthus, Torbjörn "Valross" Andersson, Paul Brown, Sean Carver, Anna Dumitriu, Leakim Eventhen, Tom Farrell, Snigdho Ganguly, Andrés Garcia, Ann-Sofi Green, David "Diddly" Guest, Michael Kamandulis, Kevin Kern, Kasia Makucewicz, Roy Mikaelsson, Jack Milner, Bhargav Mitra, Celestine Ndifor, Emmanuel Otchere, Andrea Pavan, Steve Reber, Henry Reid, Joseph Resovsky, Brenda Risch, Warren Rochelle, Enoch Sackey, Filippo Saglimbeni, Arif Said, Petr Sidlo, Florian Siegmund, Peter Smith, Mog Stapleton, Albin Sundin, Aron Vallinder, Greg "Weathercarrot" Walter, Paul Wennekes, Jamie Wilson, Tom Wright, and Tom Ziemke.

Preface

The bulk of chapters Six and Seven has been published in modified form in (Parthemore and Morse, 2010). That paper was produced as an integral part of the present thesis research. The content of these chapters is solely my own. Portions of chapters One and Five are part of another (single-author) paper that is being published as Parthemore (2011).

Contents

Chapter 1

The Quest for Concepts

1.1 The Nature and Limitations of Knowledge

One might be forgiven for thinking that human knowledge has expanded geometrically in recent decades: both our knowledge of the world around us and our knowledge of ourselves, as biological and intentional agents. Such high-profile philosophers as David Chalmers (2009) have written in favourable terms about a so-called *technological singularity*, an impending moment in the near future at which point the rate of growth of knowledge becomes so great that anything after that moment in time becomes impossible to predict.

This work is not about the unboundedness of human knowledge, either as quantifiable information or as knowledge *ability*: the ability to gain or to use knowledge to various ends. It is rather about mapping out the nature and (especially in Chapter Five) boundaries of that knowledge, at least when what we have in mind is the systematic, structured, learnable, reusable, frequently (though not always) articulable kind that we might call *conceptual knowledge*. The idea that such knowledge is bounded – even necessarily so – is an old if tendentious one in philosophy of science. According to such a view, science is less about establishing atemporal truths (if at all) and more about putting forward hypotheses that one then attempts to disprove. It is an idea with a long pedigree, with such notable adherents as Albert Einstein, as the following story suggests:

> I talked for quite a while to Albert Einstein at a banker's jubilee banquet where we both felt rather out of place. In reply to my question what problem he was working on now, he said he was engaged in thinking. Giving thought to any scientific proposition almost invariably brought progress with it. *For, without exception, every scientific proposition was wrong.* That was due to human inadequacy of thought and inability to comprehend nature, so that every abstract formulation about it was always inconsistent somewhere. Therefore every time he checked a scientific proposition his previous acceptance of it broke down and led to a new, more precise formulation. This was again inconsistent in some respects and consequently resulted in fresh formulations, and so on indefinitely. – from the diaries of Count Kessler, quoted in (Stachel, 1982, p. 96), *emphasis added*

1.2 What is a Concept?

> . . . People who start in the traditional way by asking 'What are concepts?' generally hold to a traditional metaphysics according to which a concept is a kind of mental particular (Fodor, 1998, p. 3).

On the one hand, answering the question "what is a concept?" may seem almost trivially obvious. Concepts are ideas – but, on reflection, not just any ideas, for concepts are meant to be at least fairly stable across time, and some ideas are, indeed, fleeting. Concepts are ideas with structure – but not just any structure, for the structure should be not just enduring but reliable. Concepts are structured ideas we may express through language – but not only through language, for concepts can also be communicated through gestures and pictures. Sometimes they may even be things we do not (if we lack the necessary words/gestures/pictures, possibly cannot) express at all.

Perhaps part of the difficulty – in saying what concepts are or what it is for something to be a concept – is that, in doing so, we are at the same time (as we have been all along, at least since we were children) employing concepts: particular concepts that might seem to presuppose things about the general nature of concepts. Their appearance is not just useful but obligatory. We cannot simply push them aside; they are not just the vehicle for the definition but inextricably part of it, violating our usual strict dichotomy between the thing being defined and the definition. Even as children, we are taught that it is a mistake to make reference to the thing being defined within the definition. Of course such self-reference is most obvious when it is most explicit: e.g., a dictionary definition for *graciousness* as "the property of being gracious"; but if self-reference is a problem, it is no less one for being merely implicit.

A *theory of concepts* is, for my purposes, any philosophical attempt to approach the question "what is a concept?" within an at least roughly formal framework – one that is, ideally, open to empirical testing. Concept studies is a recognizable sub-field within philosophy of mind. Over the years many theories of concepts have been put forward, some of which I will look at in Chapter Two; likewise philosophers have made many attempts to sum up, in a phrase or two, what a concept is, or what concepts are. Here is a representative selection:

- ". . . Mental particulars; specifically, they satisfy whatever ontological conditions have to be met by things that function as mental causes and effects" (Fodor, 1998, p. 23).
- ". . . Like the scale models that stand in for objects during courtroom reenactments. They allow us to reexperience past events or anticipate future events" (Prinz, 2004, p. 150).
- ". . . Complex general ideas, combining various characteristics and features" (Torey, 2009, p. 20).
- ". . . A mental representation that contains knowledge about an object or class of objects that serves to pick out or point to the object or class of objects that are characteristically associated with the concept. . . . The idea of an object is broadly defined to mean any entity or phenomenon (or classes) that can be characterized according to stored or directly perceived knowledge about the entity or phenomenon (or relevant classes)" (Hemeren, 2008, p. 15).
- ". . . A *perceived regularity (or pattern) in events or objects, or records of events or objects, designated by label* (Novak and Canas, 2008).

Over the chapters that follow, I will attempt to answer the question "what is a concept?" through a succession of working definitions and to put forward a theory of concepts that best addresses the various needs I will identify. At no point will those "working definitions" be intended as more than useful approximations. Indeed, rather than attempting to find one single answer, I will identify several (a philosophical approach known as *pragmatism*), although these will be shown to line up, more or less, on either side of one major divide. Here is a working definition to get us started:

> A synchronized relation, of some kind, between a mental agent and an experienced environment that includes that agent[a].

[a] *Cf.* (Aisbett and Gibbon, 2001, p. 190): "Representations are sometimes defined to be substates of a cognitive system that support the system's purposeful interaction with its environment. . . . For the system's interactions with its environment to be non-accidental, the state of the external environment must be able to affect the state of the system".

Let me offer two immediate caveats. One is always a product of one's background, and the questions one asks are inevitably shaped by both one's immediate environment and one's background. They are shaped, as well, by the intended context of application for the answers that one finds. In my case, I have a previous degree in artificial intelligence (AI) and "knowledge-based systems". I am first a cognitive scientist and only second a philosopher. I am motivated by broader questions about how one might do cognitive science differently. A psychologist or an engineer, or a theologian, would approach these questions, and apply their answers, quite differently.

This introductory chapter is meant to be, as far as possible, non-technical. Some of what follows will get fairly technical, but my hope is that any reasonably educated layperson will get something meaningful from even the most technical of sections. My motivation is two-fold: first, I believe that any well-written philosophical work *should* be accessible in this way. Second, my suspicion is that an overly technical analysis of the domain risks missing its intended target entirely: ironically, by over-conceptualizing it. There is much to be said for starting with, and never straying all too far away from, our everyday folk understandings[1] of such things as concepts.

1.3 Why the Question Matters

> . . . The heart of a cognitive science is its theory of concepts. And I think that the theory of concepts that cognitive science has classically assumed is in a certain way seriously mistaken (Fodor, 1998, p. vii).

At first glance, my chosen subject area might seem obscure even to many philosophers outside philosophy of mind, never mind anyone outside philosophy. Unlike other areas such as ethics or political philosophy, it may not be immediately obvious what hangs off philosophy of concepts. Concepts seem at the same time both too broad and too basic. In any case, why should it matter what they are, so long as we can use them appropriately?

Two answers concern me here, one more philosophical, the other more directly practical. First, a better understanding of the nature of structured thought is an important step to a better un-

[1] By "folk understandings" I mean terms as understood in common parlance, as opposed to how they may be philosophically re-defined.

derstanding of ourselves as *intentional agents*: (self-)conscious entities acting within the space of reasons; and along with that, a better understanding of our cognitive limitations.

Second, an improved understanding of concepts has practical consequences for both cognitive science and artificial intelligence (AI). Although in the pages that follow I will have many opportunities to disagree with Jerry Fodor, nonetheless on the matter of the quote that opens this section, I am in full agreement. Indeed, more than anything else, this is the idea that inspired the present work. Concepts are fundamental to the structure of the reasoning mind: in a real sense, they *are* that structure, or our best approximation to it. Any science that purports to be a science of the mind cannot help but address them.

Roughly, I will take cognitive science and AI to be two sides of the same scientific coin: the one more theoretically oriented, the other intending to be more practical. AI puts cognitive science theory into practice by, in one way or another, building artefacts. Sometimes the artefacts are computer programs; sometimes they are (virtual or actual) robots. By one tradition – what one might call *weak AI* – AI is about modeling aspects of (human) intelligence and enabling artefacts to do what previously only humans could do. By another – what one might call *strong AI* – AI is about creating independently intelligent, even self-conscious, artefacts[2].

AI has attempted to capture conceptual knowledge in a multitude of ways, most often through what have been called *knowledge representation* (KR) *formalisms*[3]. At one time – particularly in the early 1990s – "knowledge representation" was one of the most talked about phrases in cognitive science and AI, enjoying something of the trendiness of "enactivism" today.

The choice of KR formalism is typically dictated by such practical concerns as the requirements of the project, the deadlines for producing a working application, the programming background of the researchers, and so on. Such concerns are of course valid. At the same time, the choice of formalism will have significant consequences for everything that follows. Meanwhile the question of whether the knowledge is being structured in anything like the way a human agent would structure that knowledge – indeed, whether that is even important – is rarely addressed.

Consider this simple program written in Prolog[4], a programming language based on Horn-clause logic[5], which is meant to capture some knowledge about James' family tree. Anything in upper case is a variable; anything in lower case is a constant:

```
parent (john, james).
parent (janice, james).
male (john).
male (james).
female (janice).
father (X, Y) :- male (X), parent (X, Y).
mother (X, Y) :- female (X), parent (X, Y).
```

[2] I am, of course, vastly simplifying matters. For an excellent discussion of some of the finer distinctions one can make, see (Sloman, 1985).

[3] . . . Not always. As I shall discuss in Chapter Two, a substantial number of researchers do not consider concepts to be representations either at all or in the first instance, and they have inspired a competing movement within AI, according to which knowledge should not be represented (at least not explicitly) but instead arise "naturally" out of systematic associations arising from the artefact's interaction with its environment.

[4] It is, in fact, a reproduction of the first computer program I ever wrote.

[5] A horn clause is a disjunction of literals (either atomic symbols or their negation) with at most one positive literal: e.g., $\neg p \vee \neg q \vee r \vee \neg s$ ("not p or not q or r or not s"). Such clauses can usefully be re-written as implications like this: $r \leftarrow (p \wedge q \wedge s)$ ("r if p and q and s").

```
son (X, Y) :- male (X), parent (Y, X).
daughter (X, Y) :- female (X), parent (Y, X).
```

The program contains some "facts" that can be interpreted by an observer (not by the program itself!) as e.g. "John is the parent of James" and "John is male", plus some "rules" according to which a number of additional, implied facts can be derived. That is to say, we know e.g. that John is the father of James, because John is male and John is the parent of James. Likewise Janice is James' mother. The more "facts" and "rules", the more implied "facts" can be derived.

Prolog is very good at capturing the many implied relations in a family tree. Its ability to tease out the logical structure in language made it an early popular choice for *natural-language processing*. In many ways, it captures deductive propositional reasoning beautifully. Some AI researchers, notably the Cyc Project's Doug Lenat, think that human conceptual knowledge really is structured in something like this propositional style (see Section 4.5.1).

> Keep explaining the meaning of these terms... and slowly this converges after writing millions and millions of these assertions into a set of axioms that have only one model, namely the real world. And finally, when you have written enough, you can believe that the conclusions that would deductively come from all these assertions would be the same conclusions you would believe about things in the real world (Lenat, 2006).

Many researchers would not be willing to make any such commitment. The problem is, it's often not clear where their commitments lie. Bottom line: *ceteris paribus*, all practical applications depend on theoretical foundations; and, when it comes to AI and cognitive science, those foundations perforce include some, often never more than implicit, theory of concepts. As in other domains, much can be gained by making the implicit explicit and reflecting on the match between theory and application.

1.4 Finding a Neutral Metaphor

A useful metaphor – one I shall draw upon a number of times in the pages to follow – is of concepts as building blocks. Of course, as with any metaphor, one should take care not to confuse the metaphor with the so-called literal meaning (though, as I will suggest in chapters Five and Six, the distinction between metaphorical and literal meaning may be more of a continuum than a clearly drawn line). As with any good metaphor, it depends both on the similarities and the differences.

Like the child's toy, concepts are (or seem to be) similarly structured, and can be placed or piled together into an endless variety of complex structures. Unlike the toy, concepts even far removed from one another may still be somehow importantly connected or associated. That is to say, if one locates concepts in a kind of *conceptual space* – another metaphor I will make much reference to – then one finds that they depend on both local and distal connections with each other.

Of course no metaphor is truly neutral. Thinking of concepts as building blocks implies not thinking of them in other ways. Still, both conceptual building blocks and conceptual spaces will prove to be – borrowing another, well-traded metaphor – powerful *intuition pumps.*

If concepts are building blocks, then a new set of building block primitives may lead to new ways of doing cognitive science research and a new generation of AI applications. If the blocks bear some close relation to our own mental structures, then anything built from them can be expected to *feel* like our own, "personal" conceptual knowledge. Conversely, if what is structured from them feels sufficiently natural, that is (indirect) reason to think that the blocks bear some important relation to our own mental structures. The goal of the present work is to make some substantial progress toward saying what a set of building block primitives might be like, how they might be placed or piled together (by what implicit rules), and how they might be used and put to use: both theoretically and practically.

1.5 Setting the Boundaries

One of the key concepts in this work will be the concept of boundary, one of the key conclusions that *absolute fixed* boundaries of nearly any kind are conceptually problematic. By "boundary" I mean, for purposes of this work, a categorical dividing line such that specific instances of entities for whom the boundary is relevant are meant to fall to one side of the boundary or the other. Any that fall directly on the boundary indicate a problem at the least with where the boundary has been drawn, if not with the boundary itself. (Of course some boundaries are inherently vague, such as the one between baldness and non-baldness. My issue, for the most part, will be with boundaries that are meant to be non-vague.) How one tends to look at boundaries depends very much, I will argue, on one's metaphysics. I believe that issues of boundary deserve a great deal more attention than they have mostly been given to date.

1.5.1 Metaphysical Starting Points[6]

In Section 5.2.3.2, I discuss the relevance (to theories of concepts) of the so-called *extended mind hypothesis*, proposed by Andy Clark and David Chalmers, which Clark describes as follows:

> Proponents of the extended mind story hold that even quite familiar human mental states (e.g., states of believing that so and so) can be realized, in part, by structures and processes located outside the human head. Such claims go far beyond the important but far less challenging assertion that human cognizing leans heavily on various forms of external scaffolding and support. Instead, they paint mind itself (or better, the physical machinery that realizes some of our cognitive processes and mental states) as, under humanly attainable conditions, extending beyond the bounds of skin and skull (Clark, 2008, p. 76).

One of the principal critics of the extended-mind hypothesis is Robert Rupert, who has recently devoted a book to the topic (2009a). Rupert puts a lot of weight on the word "literal": the word is "literally" peppered throughout his book. In a typical passage he describes the extended view as "the view that human cognition – to some substantial degree – literally includes elements beyond the boundary of the human organism" (Rupert, 2009a, p. 3). The implication, I assume, is that Clark and Chalmers are not "merely" speaking metaphorically: they "really" mean that. Such a crisp literal/metaphorical distinction – whether with one's language or one's concepts – assumes some form of realist metaphysics, as is clear from many places in Rupert's writings (though nowhere

[6] Much of the content of this section appears in (Parthemore, 2011).

have I seen it stated so baldly)[7]. The literal meaning is the fact of the matter that realism aims to deliver. But there is nothing about physicalism (or materialism or naturalism), so far as I am aware, that entails realism. On this point Clark is holding his cards in his hands, while Chalmers is, so far as I can tell, best understood as both a physicalist (he rejects substance dualism) and an anti-realist (he takes experience as foundational: i.e., something that must be assumed from the beginning) (see e.g. (Chalmers, 1996)). Contrast Rupert with Peter Gärdenfors, whose anti-realist leanings inform his theories about concepts: "The upshot is that [in conceptual spaces theory] there is no sharp distinction between literal meaning and metaphor" (Gärdenfors, 2004, p. 187). Gärdenfors' conceptual spaces theory will play a key role in the chapters to follow.

Realism I take to be the metaphysical assumption either (*per* direct realism) that the apparent transparency of the world should, in most instances at least, be taken at face value; or (*per* indirect realism) that if the apparent transparency cannot be taken at face value, it can, at least, be logically reconstructed. In either case, science talks about the world in a perspective-free (or essentially perspective-free) way. Meanwhile anti-realism is the position that, while the fully mind-independent world is conceded logically to exist, one cannot, as a matter of principle, say anything about it; or that the world we experience is always in some way touched by mind.

Let me be clear: anti-realism is not the perspective that world *is* mind; that would be idealism. Neither does anti-realism allow one to believe whatever one likes about the world: if the world constantly outruns our conceptual understanding of it, at the same time it constantly and forcibly constrains that understanding, sometimes on pain of injury or death.

Anti-realism, pragmatism, and pluralism go hand in hand, where pragmatism is taken as the position that there need be, in most instances at least (and perhaps all of any import), no single fact of the matter. Pragmatism can even tolerate apparent contradictions, so long as they are qualified by perspective: e.g., p from one perspective, $\tilde{}p$ from another. So long as one does not try to hold both perspectives at once – i.e., make them part of a single perspective – there is no contradiction in possessing both of them.

In this light, Rupert's statement that "even if one is inclined toward pluralism, an extended and an embedded model cannot both be true of a single cognitive process – else there is a single cognitive system that both extends beyond the boundary of the organism and does not" (Rupert, 2009a, p. 9) is, on the face of it, simply wrong. Pluralism and pragmatism in no way guarantee that our conflicting perspectives can necessarily be reconciled[8].

Like Gärdenfors, I am an anti-realist by inclination, and my anti-realist intuitions will, doubtless, inform much of this work. I will not, however, argue either for the correctness of anti-realism nor the incorrectness of realism – nor need I do so. It is enough to allow a modest anti-realism as a plausible position for sake of argument. If some form of anti-realism *should happen* to be true,

[7] As one of the reviewers for (Parthemore, 2011) commented in objecting to my portrayal of Rupert as a (direct) realist, Rupert is a representationalist. So far as I can see, there need not be any conflict between being a representationalist and being a direct realist, if the representations are suitably transparent: convenient mental shorthands, as it were. However, for my purposes, it is enough to note that there is *no* conflict between representationalism (in its many forms) and indirect realism, and all I need to show is that Rupert is assuming *some* form of realist metaphysics.

[8] Indeed, Rupert elsewhere might seem to allow this, given he acknowledges the likelihood that "different representations are used depending on context and thus that the subject represents what might normally be characterized as the same thing in different ways depending on context" (2009a, p. 196). I see nothing in that passage to indicate either that one of the representations will be the "right" one or that the representations may not be in seeming conflict with each other.

then intuitions, like perspectives, cannot simply be set aside – *pace* Rupert (2009a, pp. 20, 32, 45); they will play an unavoidable and substantive role in the theory. This raises the first of three critical distinctions I will be making in this work: between theories of concepts as informed by a realist versus anti-realist perspective.

1.5.2 Metaphysical Premises and Boundaries

Metaphysical premises become clearest when one looks at the concept of boundary[9]. After all, the extended mind debate at heart is about where one should draw the boundary between mind and world, and whether that boundary is fixed at the physical boundary of skin and skull. For all of the importance (rightly) placed on this boundary, one might expect there to be more attention paid not just to locating it correctly but determining its nature. Is the boundary "really" real, or is it something we construct (and can move)?

Notions of "inside" and "outside" are not always so clear even when it comes to physical volumes, whose boundaries seem at first blush to be so clear. Consider cell boundaries, as clear of a boundary as one is likely to find. Any effective cell membrane must be permeable: a continuity to match the discontinuity. The problem is: at what precise point does a molecule pass from "outside" to "inside"? The closer one examines the cell boundary, the harder the answer becomes. The answer is only clear if one observes from a sufficiently detached perspective.

Boundaries at the level of multicellular organisms only become more difficult. Is my bodily boundary at my epidermis (layer of dead skin cells) or my dermis (live cells)? It depends upon the context in which you're asking. Likewise my body is, topologically speaking, torus shaped. Normally I think of my digestive tract as "inside", but from some perspectives it is "outside".

What of the bacteria living in my gut, who depend on me for their existence, and whom I likewise depend on for mine? Are they inside or out? Are they part of me or not? I am reliably informed that several kilograms of my body weight consist of single-celled organisms: some symbiotic, some neutral, some parasitic. When I weigh myself in the morning, I do not mentally subtract them.

The realist, of whatever persuasion, need not, of course, be bothered about any of this. *Prima facie,* it is enough for her to say with respect to the extended mind debate that e.g. the boundary between mind and world is *roughly* at the physical boundary of skin and skull – or is it?

The difficulty (or, for the extended mind enthusiast, the opportunity) lies with how rough is "roughly", and in particular with the way Rupert (along with Adams and Aizawa) moves seamlessly from the boundary of the organism *as a biological agent* to the boundary of the organism *as a cognitive agent*. "Internal" and "external" are attributes of physical objects (or collections of such objects), and it is not immediately clear whether a cognitive agent is that sort of thing. Yet Adams and Aizawa write, "To ask about the bounds of cognition is to ask what portions of spacetime contain cognitive processing.... It is to ask about the physical substrate of cognition" (2008, p. 16).

[9] I talk here about fuzzy boundaries; but there is, of course, a comparison to be made with *fuzzy logic* (or *fuzzy set theory*) as originally proposed by L.A. Zadeh (1965). Both discussions involve reasoning with uncertainty. In the case of fuzzy logic, the uncertain boundary is between one definite conclusion and another, mutually exclusive, definite conclusion.

This is, perhaps, not a problem, provided one sees mind reducing to brain *per* an eliminativist account (such as the Churchlands offer), or mind emerging from brain in a way that is either immediately transparent or reconstructibly so. (The latter route is, I think, the one that Rupert wants to take: he seems ready to allow that mind *could* just be a functional description with no immediate physical translation.) Either mental boundaries "just are" physical boundaries, or they map straightforwardly to them.

Reconstructible in principle, however, need not mean reconstructible in practice; and herein lies the fruitful middle ground between anti-realist and realist perspectives: without that reconstructibility, mental boundaries look woefully unclear. At the same time, a clear and at least relatively fixed mental boundary is essential to Rupert's arguments[10].

In keeping with much if not most of the literature in this field, Rupert talks about representations without defining what they are: the assumption is that the definition is already understood and agreed upon. (I will devote considerable attention to defining representations in Section 2.6.) Unfortunately, proceeding to label some representations as "internal representations", as Rupert does, does not help unless the application of "internal" in the mental domain is *also* already understood and agreed upon. Rupert's offer of a "systems-based criterion" (2009a, ch. 3) is, on its own, no help recapturing a clear sense of boundary unless that criterion is assuming the very physical translation that is meant to be derived – no help, certainly, if one gives any weight to concerns like this one from Clark:

> Nontrivial causal spread... occurs whenever something we might have expected to be achieved by a certain well-demarcated system turns out to involve the exploitation of more far-flung factors and forces (2008, p. 7).

One school of thought actively trying to occupy the middle ground is enactivism, a school of philosophy that views cognition as spanning brain, body and environment. Enactivism will be mentioned at several points over the following chapters and discussed in some detail in Section 6.1.2.

1.6 Conventions

At some points in the text, I will refer to "the concept of X" or "the concept of an X". At many points, however, I will abbreviate this by putting the concept in boldface: i.e., **DOG** should be understood to mean "the concept of a dog", while **MY DOG FELLA** should be understood to mean "the concept of a particular dog whose name is Fella".

1.7 Structure of the Book

Chapter Two addresses the basic nature of concepts and of the theories about them. Historically, concepts were most often understood either as dictionary-like definitions or as (actual or metaphorical) "images in the mind". Influences of both traditions remain. Contemporary accounts differ over whether concepts should be understood, entirely or at least in the first instance, as

[10] Contrast this with Clark's flexible notion of the same boundary, when applied to what he terms "profoundly embodied agents": "Such agents are able constantly to negotiate and renegotiate the agent-world boundary itself. Although our own capacity for such renegotiation is, I believe, vastly underappreciated, it really should come as no great surprise, given the facts of biological bodily growth and change" (2008, p. 34).

representations (often referred to as *mental representations*) or as abilities (e.g., abilities to form such representations). In order to decide between them, I raise the second of my three critical distinctions: between concepts as we reflect on them, and concepts as we possess and employ them non-reflectively. The concept of representation (and the related concept of symbol) require considerable examination. Roughly, I will take a representation to be any *something* used by *someone* to stand in place of *something else* for *someone else* (or for oneself). Meanwhile if concepts are abilities, then to have a concept is to show a certain competency: e.g., to possess the concept of a dog is (nothing more than) the ability to respond appropriately (under normal circumstances) to dogs *as* dogs (and not as, say, cats) – as well as to be able to read, write and talk about them in a way consistent with societal standards.

Chapter Three has two principal goals. First, it sets out a list of properties of concepts and sorts them into the essential ones versus those that are commonly associated with concepts but not essential to them. Even though it is far from universally accepted, nonetheless Gareth Evans' Generality Constraint (1982, p. 101 ff.), with its implied claims to the systematicity and productivity of thought, is taken as a baseline. I introduce the longstanding debate over whether or not concepts (or conceptual abilities) are (either by stipulation or as a matter of empirical analysis) co-extensive with language (or linguistic abilities). I conclude that concepts not only need not be articulable by the agent possessing them; they need not even be introspectible by that agent. (Indeed, as Chapter Four will conclude, certain concepts logically cannot be.) In doing so I introduce the last of my three critical distinctions, between the public and the private aspects of concepts. The latter part of Chapter Three then sets out the necessary conditions on a successful theory of concepts, taking inspiration from similar attempts by Jerry Fodor (1998), Jesse Prinz (2004), and Eric Margolis and Stephen Laurence (1999; 2002).

Chapter Four places both concepts and the theories about them into a context of use and of the agents who possess and employ them. Its guiding principle is that any given concept does not exist in a vacuum but always attaches to two things: one or more conceptual agents possessing and employing it, and one or more referents that it is about; in slogan form, concepts are always somebody's concepts *of*. One of its central conclusions is that categorization does not map categories of concepts to corresponding categories in the world; rather, categorization imposes conceptual structure onto the world. Concepts are distinguished first- from second- and higher-order, physical from mental, (relatively) static from (explicitly) dynamic, laying the groundwork for three of the four axes required by the *unified conceptual space theory* (Chapter Six). While Chapter Two placed theories of concepts into a historical context, Chapter Four, in analogous fashion, places concepts into an evolutionary perspective, the better to argue that concept possession is not exclusive to human animals. The chapter concludes by putting theories of concepts into a context of the questions they are asking and the applications to which the answers will be applied. Two cautionary tales are offered. A single theory of concepts that works across *all* contexts is probably unachievable and in any case undesirable.

There are places our conceptual understanding cannot take us; past a certain point, it breaks down. So Chapter Five addresses the inherent limits of conceptual abilities: both the "soft" ones (which we can perceive in a general way, and so explore) and the "hard" ones (which we can only conclude to exist, outside our ability to perceive). It presents what I call the Hard Problem of Concepts, a translation or re-interpretation of David Chalmers' Hard Problem of Consciousness (1996): the difficulty, or impossibility, for the conceptual agent to separate out her concepts from

experience that, in some basic ways, presupposes them. Conceptually structured experience is, ultimately, neither an explanation *nor an explanandum* but rather a starting condition. Concepts are *necessary fictions*, not so much true as true enough, simplifying in pursuit of understanding, creating what are, for the most part, *innocent inconsistencies* between what they purport to and what they actually describe. Self-referential paradoxes arise when we press against our conceptual boundaries. I recapitulate and criticize a famous argument from Roger Penrose that, if correct, would seriously undermine these conclusions. If the "price" of the chapter is an acknowledgment of our limitations, the "prize" is a powerful tool for toggling back and forth between competing perspectives.

Chapter Six presents Gärdenfors' *conceptual spaces theory* of concepts (2004) as the contemporary theory best able to meet the conclusions of the previous chapters and bridge between the useful insights of other, competing theories. (After all, if inconsistency cannot be eliminated entirely, one still wants a principled means of reconciling inconsistent accounts in a way that does justice to the strengths and weaknesses of each, without leaving one open to a charge of "anything goes".) It acknowledges the limitations in conceptual spaces theory – in particular, the lack of detail (which Gärdenfors acknowledges) – then offers a proposed set of extensions to it that collectively I call the *unified conceptual space theory*. This attempts to bring together all the many heterogeneously structured conceptual spaces that Gärdenfors describes, within a single space of spaces oriented along four axes, divided into well-behaved and, for the most part, convexly shaped sub-regions with both local and distal connections within that space. These sub-regions fall into one of three highest-level subregions, defined by their location along two of the axes and by the sorts of distal connections they countenance. Any concept may be given either of two descriptions: one as a location along the four axes, the other as a set of distal connections.

Chapter Seven offers the powerful notion of *concepts as expectations*, structuring, guiding and simplifying our interaction with our environment and with ourselves, allowing us to interact with a conceptually integrated whole rather than non-conceptually non-integrated parts. Its primary goal is applying the unified conceptual space theory to a framework for the co-emergence of concepts and experience. Concepts are grounded in a combination of sensorimotor and somatosensory interactions and a pattern recognition process that recognizes certain perceptual regularities as salient and remembers them. Experience, meanwhile, is increasingly, as we develop as conceptual agents, grounded in concepts, so that its non-conceptual content is overshadowed or forgotten; only when conceptual expectations break down do we "take a closer look". Concept acquisition (experience giving rise to concepts) and concept application (concepts structuring or creating experience) are to be understood logically as dynamically coupled processes but best approached conceptually within a circularly causal framework where each is cause and effect of the other. By an inversion of perspective, the one process becomes the other: the underlying mechanism is neutral as to whether one is acquiring concepts or applying them.

Chapter Eight changes tracks by adopting more of a traditional cognitive science-oriented rather than philosophical tone. It describes the prototype of a software tool for assisting users in creating an externalized model of some conceptual domain or another. The tool takes inspiration from so-called mind-mapping software but departs fundamentally from it both in its style of interface and in its reliance on a clearly articulated theory of concepts. That theory is, indeed, a more-or-less direct translation of much of the unified conceptual space theory, allowing it to be mapped out along all four dimensions and navigated along three. At the same time, the intended user should

require only the most basic explanation of the underlying theory in order to make full use of the tool; the more intuitive its operation, the less explanation will be needed. Even though the tool is, on the face of it, the essence of unembedded, disembodied systems, nonetheless, these things should not be seen as absent but rather offloaded onto the user, who provides the embeddedness and embodiment that the tool on its own lacks. The implementation serves two purposes: it delivers on the promise of putting theory into practice, and it provides a concrete visualization of that theory. Already the software tool has motivated revisions in the theory.

Chapter Nine compares what I set out to do with what I accomplished. It looks back over the preceding chapters to try to capture the bigger picture, to draw some conclusions, and offer a few self-criticisms. At the same time it looks forward, offering the broadest outlines of where I would like to take these ideas next.

Chapter 2

Toward an Ontology of Concepts

There are many things to be said in the chapters to come about the nature of concepts: their core properties (sections 3.1 and 3.2), their basic sub-types (sections 4.1 and 4.2), their ontogenetic (sections 3.4.2 and 7.2) and phylogenetic (Section 4.4) development, and the uses to which both they (Section 4.3) and the theories about them (Section 4.5) are put. For all their seeming diversity, a number of common patterns will emerge. But first, there is a need to get clear – or as clear as possible – about the basic ontology. Setting issues of properties and context aside: if a cat is a mammal and a democracy is a form of government, what sort of thing is a concept of a cat or a concept of a democracy, or for that matter, the concept of a concept itself? What *ontology* must we (assuming a single correct one), or should we (assuming the possibility of several valid ones), be committed to? Are concepts, as Fodor concludes they must be, "mental particulars", or is the notion of a mental particular an empty one? Are they causes and effects, or are they epiphenomenal?

An ontology is a philosophical specification of what-ness. Ontologies are interested not in *how* we come to know something (the realm of epistemology) but *what* it is that we come to know: i.e., its essential nature. Here, questions of metaphysical prejudice intrude: for realists, that essential nature will be something mind independent and determinate, such as physical symbols in the brain (*per* the *physical symbol system* hypothesis (Newell, 1980)); while for idealists, that essential nature will be fully mind dependent and possibly indeterminate, as shiftable as the flow of our thoughts themselves. Anti-realists, on the other hand, both assume a mind-independent reality and place it beyond either understanding or experience, even as it is continuously and intimately constraining both. They will typically be looking for an answer that assigns concepts both a physical and a mental nature, neither reducible in practice to the other[1]. At the same time, like the pragmatists with whom they have much in common, they will be disinclined to look for a single correct answer – a unitary fact of the matter – for all but the most trivial of questions.

Questions of metaphysics will come to a head in Section 5.2. While remaining largely neutral on its metaphysics, this chapter will conclude, in pragmatist fashion, that a single ontology for concepts will not be sufficient, and that two contrasting ontologies are required.

[1] If they were irreducible in principle, this would be Cartesian *substance dualism*.

2.1 Folk Foundations

> When one says "dog", or "yawl", or "junta", there is the strong impression that discrete ideas apply to each of these words. Cognitive science, following common sense, calls such ideas "concepts". Typically at least, we reflective common folk think of a word as the expression of a concept, of concepts as the constituents of larger mental units, such as thoughts, and thus of concepts as central to our mental life (Keil and Wilson, 2000, p. 308).

The folk understanding of concepts, as gleaned from e.g. various dictionary entries, might usefully be summarized as an abstract idea or general understanding arrived at from specific instances or experiences[2]. Consider that one's concept of Fella (a dog) is derived from various Fella encounters, while one's concept of dog is derived from encounters with Fella and other dogs (or representations or descriptions thereof). One's concept of animal is likewise derived from encounters with dogs and other animals. No order of concept acquisition should be taken to be implied here: one might well have a concept of dog before one has a concept of Fella or *vice versa*. A particular model of concept acquisition however *is* implied – one that may or may not bear up to empirical scrutiny[3].

Figure 2.1: *Matryoshka* dolls are unlike folk concepts, in that any folk concept may contain *many* concepts "inside" itself. (Photo downloaded from Wikimedia Commons: http://commons.wikimedia.org/.)

The folk understanding of concepts assumes a hierarchy. Concepts nest inside one another not unlike Russian *matryoshka* dolls (see Figure 2.1). Fella, as it happens, is a boxer; a boxer is a type

[2]Indeed, I arrived at that definition by generalizing over a number of dictionary entries.

[3]It is worth noting in passing that what goes for "object" concepts like **FELLA** goes *prima facie* for "action" concepts like **FETCH** as well: i.e., one's concept of fetching is, on this account, derived from experiences/observations/accounts of both various types and instances of fetching. More will be said about categories of concepts in Chapter Four.

of dog (all boxers are dogs); a dog is a type of mammal (all dogs are mammals); a mammal is a type of animal (all mammals are animals). There are *prima facie* no exceptions: a boxer is never not a dog; a dog is never not a mammal. Because a dog is a mammal, a dog must be an animal, too. Because a dog is a mammal and all female mammals[4] bear live young, a female dog bears live young. I will say more on hierarchy and inheritance in Section 6.2.1.

This hierarchical organization makes concepts roughly synonymous with categories, which can, themselves, be members of other categories: **DOG** is a category containing various dog breeds; it is a member of the category **MAMMAL**. At first blush, **MY DOG FELLA** is an exception: seemingly it is *only* a member of a category and not a category itself. The hierarchy, which began with **ANIMALS** or **ORGANISMS** or, more broadly yet with **THINGS**, bottoms out; the *matryoshka* doll reveals a final doll that is different, because it does *not* open to reveal another doll.

However, things may not be so straightforward since, even with **MY DOG FELLA**, or indeed any concept of a particular one can name, the pattern of generalization over specific instances remains. **MY DOG FELLA** does not refer only to **MY DOG FELLA AT 14:03 YESTERDAY** or to **MY DOG FELLA THE LAST TIME HE WET THE CARPET** but to **MY DOG FELLA** on these and all other occasions, and potentially many occasions yet to come. In this way, **MY DOG FELLA** is itself a category, bringing together all the Fella-directed or -related experiences and thoughts. This is, in fact, the approach I will take in putting forward my own theory of concepts beginning in Chapter Six; see in particular Section 6.1, where I will follow Gärdenfors in rejecting any substantive class/instance distinction.

The Keil and Wilson quote above offers a few further points. Concepts are or are meant to be discrete ideas, individuable and hence distinguishable one from another. Folk notions of concepts are closely associated with folk notions of language. Words give expression to our concepts, so that much if not most of the time concepts line up neatly with corresponding words or phrases and *vice versa*: the word "dog" expresses the concept **DOG**[5]. Most importantly, concepts are the basic building blocks of thoughts and so of the rich tapestry of our mental lives.

Some philosophers, notably the Churchlands (see e.g. (1981)), would deny any merit to folk understandings of concepts whatsoever: such understandings, they would say, are too muddled and too dependent on questionable presuppositions. In response, and in the spirit of Wittgenstein or the plain language philosophers, I would suggest that it is at least equally plausible that the folk understandings are the clearest and that over-analysis makes matters worse, not better. "The aspects of things that are most important for us are hidden because of their simplicity and familiarity" (Wittgenstein, 2001, § 129), or more baldly: "When we do philosophy we are like savages, primitive people, who hear the expressions of civilized men, put a false interpretation on them, and then draw the queerest conclusions from it" (Wittgenstein, 2001, § 194).

The idea that philosophers of concepts can completely discard our folk understandings and start over with a completely new approach incommensurate with those folk understandings is difficult to motivate. *Prima facie,* and in absence of evidence to the contrary, philosophers of concepts are as much bound by their folk understandings as anybody else. What philosophers *can* do is start

[4] . . . With the exception of the platypus.

[5] In philosophical terms, *lexical concepts* relate to individual words and *complex* or *phrasal concepts* to phrases: e.g., **BROWN COW**.

with those folk understandings and attempt to refine them, to put flesh onto their bones. In any case, the burden is on anyone who would deny any value to folk understandings of concepts to offer some account of how they are meant to be set aside.

2.2 Philosophical Questions, Philosophical Boundaries

> The whole idea of the mind explaining itself is a logical contradiction (Hayek, 1999, p. 192).
>
> If you work on your mind with your mind, how can you avoid an immense confusion (Watts, 1957, p. 94)?

Folk understandings of concepts may be sufficient for the layperson's everyday usage and reflection, but as the basis for a philosophical theory of concepts they seem grossly inadequate.

- What does an "abstract idea" or a "general understanding" amount to?
- What is the empirical evidence for concept acquisition?
- Is the hierarchical organization of concepts real or illusory?
- Are concepts the same as categories and, if not, how do they differ?
- Are "object" concepts and "action" concepts instances of a common genus or not?

A far more general and philosophically challenging question is lurking, however. How can one reflect upon or discuss the nature of concepts without, in a non-trivial way, presupposing the very conceptual structure one is attempting to analyze? When we conceptualize self-consciously, we do so from within a pre-existing conceptual structure; concepts – not least the concept of a concept itself – are routinely if not exclusively accounted for in terms of other concepts. Pressed too far, such circular reasoning leads to confusion at best, at worst – as Hayek notes – to outright contradiction. Much more will be said about this in Section 5.2.1.2. For now, two points should be born in mind:

- A theory of concepts may be expected to try as much as possible to ground concepts in something non-conceptual and so escape that circularity. More will be said about this when I talk about Ron Chrisley's notion of non-conceptual conceptual change (Section 3.2.4).
- Even if philosophical analysis of concepts is necessarily bounded (as indeed I believe it is) and a *full* account of concepts or of mind is a contradiction, that does not mean we should not press our understanding as far as we can. It is in that light that the present work is intended.

2.3 Philosophical Traditions

Taking on board the folk understanding of concepts as abstract (mental) objects, philosophers of concepts have historically fallen into two camps as to exactly what those objects might be: definitionism and imagism. One stresses the relationship of concepts to (words of a) language (paradigmatically symbolic[6]), the other to some sort of pictures (paradigmatically non-symbolic).

[6] I will define symbols in Section 2.6.

2.3.1 Classical Definitionism

> Thus, the idea of the sun – what is it but an aggregate of those several simple ideas, bright, hot, roundish, having a constant regular motion, at a certain distance from us, and perhaps some other (Locke, 2004).

By most accounts (e.g., (Fodor, 1998; Laurence and Margolis, 1999, 2002; Prinz, 2004)), the dominant tradition until somewhere around the middle of the last century held that concepts were definitions, roughly in the style of dictionary entries. Indeed, Stephen Laurence and Eric Margolis simply call this the Classical Theory of Concepts. "Definition" can be understood more literally or more metaphorically; the words of the definition can either be words of a public language or the equivalent translation into the private "language of thought" that Fodor (1975; 2008)has famously called *mentalese*.

To be more precise, what classical definitionist accounts have in common is the assumption that concepts consist of an explicit set of *necessary* and *sufficient* conditions for their proper application (as provided by the "definition"): *if* the conditions are met (sufficient) and *only* if the conditions are met (necessary), then the concept "fits". Concepts are composed out of other concepts, down to the level of some primitive concepts, which are usually (but not always) taken to be sensorily grounded. Laurence and Margolis (1999) single out John Locke based on quotations like the one above[7]; other names that get mentioned are Plato, Descartes, and Kant.

Finding the set of necessary and sufficient conditions is easier said than done, as Wittgenstein famously showed with his example of games.

> How should we explain to someone what a game is? I imagine that we should describe games to him, and we might add: "This and similar things are called 'games"'. And do we know any more about it ourselves? Is it only other people whom we cannot tell exactly what a game is? – But this is not ignorance. We do not know the boundaries because none have been drawn.... [Someone says to me: "Shew the children a game." I teach them gaming with dice, and the other says "I didn't mean that sort of game." Must the exclusion of the game with dice have come before his mind when he gave me the order?] (Wittgenstein, 2001, § 69)

The classic example of where definitions look their best and go most wrong is "bachelor = unmarried male". At first glance, a person is a bachelor if and only if the person is male and unmarried. But in that case what does one make of the Pope or a young child or, for that matter, someone who is living together with someone outside of the legal structure of marriage? The more one examines the definition, the more qualifications, in non-monotonic style, need to be added.

The chief argument against definitionism – and the reason classical definitionists are so thin on the ground – is that, as Fodor states, "there are practically no defensible examples of definitions..." (1998, p. 45). Indeed, it is unclear whether there are any airtight definitions, even in mathematics[8].

[7] Note that the same quote could be used to portray Locke as a prototype theorist and hence a kind of imagist: see Jesse Prinz's comment below.

[8] Consider "even number": "any integer divisible by two, such that it leaves no remainder, and such that the set containing all instances of it is mutually exhaustive to the set of all odd numbers, which leave a remainder of one". Given that the set of integers is meant to be infinite, one can separate the integers into the finite and the non-finite. If it is clear whether any finite integer is even or odd, it is not so for a non-finite integer. If one objects to talking about non-finite integers, there are still uncountably large ones, such as the number of stars in the visible universe,

Lacking such fixed definitions, definitionism ends up looking like an *ad hoc* means to an end; as Fodor wryly notes, "It's a sad truth about definitions that even their warm admirers rarely loved them for themselves alone" (Fodor, 1998, p. 69).

Definitionism has other problems. Definitions are either met or they are not; there is no sense to be made, on the classical account, of *almost* meeting them or *just* meeting them or meeting them particularly well. It is indeed true that one cannot be almost a bird or just barely a bird or very much a bird. At the same time, some birds are considered more typically bird-like than others: a robin, say, compared to an ostrich[9]. Likewise for animals that are not birds: some are considered more like a bird than others – say, a platypus (which lays eggs) or a pterodactyl or a bat (both of which fly), compared to a moose.

Category boundaries raise further problems for definitions. At first blush it may seem quite clear what is or is not a bird; but there are borderline cases – for example, with ancestors of modern birds – where even experts might debate how to classify the animal. Meanwhile, what does one make of a dead bird? Is it still a bird? It may be easier to say it's a bird if it's recently deceased than, say, if it's decomposed into a skeleton. At what point does it stop being a bird / meeting the definition of a bird?

Even the most apparently well-defined concepts can be found, in the right circumstances, to have unclear boundaries: the problem of conceptual vagueness or *fuzziness*[10]. Our concepts can navigate these vague boundaries; definitions, at least in the classical sense, cannot. Therefore, it is said, concepts cannot be definitions.

2.3.2 Classical Imagism

> . . . A *triangle* is defined to be a *plane surface comprehended by three straight lines*; by which that name is limited to denote one certain idea and no other. To which I answer, that in the definition it is not said whether the surface be great or small, black or white, nor whether the sides are long or short, equal or unequal; nor with what angles they are inclined to each other; in all which there may be great variety, and consequently there is no one settled idea which limits the signification of the word *triangle* (Berkeley, 1999, p. 18).
>
> What really comes before our mind when we *understand* a word? – Isn't it something like a picture? Can't it be a picture? (Wittgenstein, 2001, § 139)

If Laurence and Margolis, and Fodor, focus their attention on classical definitionism, then Jesse Prinz (2004) offers something close to equal press to imagism, whose roots may be said to go just as far back. By imagist accounts, concepts are – more literally or more metaphorically, depending on the concept in question – pictures (shapes, or collections of shapes) in the mind. What definitionism sought to do with necessary and sufficient conditions, imagism sought to do with similarity. The usual method of measuring similarity was resemblance: a concept "fits" if it sufficiently resembles

which is either even or odd *in principle* but cannot be determined to be even or odd in practice, which again might seem to raise a problem for the definition.

[9] When asked to name birds, people will standardly list more typical examples before less typical ones, with considerable cross-societal agreement about what counts as typical. In psychology, such results are known as *typicality effects.*

[10] The problems of definitionism with typicality effects and conceptual fuzziness are nicely discussed in (Laurence and Margolis, 1999, 2002).

its intended target. Likewise two things are conceptually related – members of some common category – to the extent they resemble each other.

Difficulties with typicality and conceptual fuzziness are, *prima facie*, resolved, since similarity, unlike necessary and sufficient conditions, does not make a binary distinction: rather, it exists along a continuum, from the completely similar (any thing may be said to be maximally similar to itself) to the completely dissimilar (i.e., no possible point of comparison). Typical members of a category bear a high degree of resemblance to each other and to the *prototype* for that category; less typical members still bear sufficient resemblance to the category prototype to count as members of the category; non-members can be seen to bear greater or lesser resemblance to the category prototype – which offers some explanation why bats are confused with birds or whales with fish.

Historically, imagism has, even more than definitionism, close ties to empiricism: the philosophical tradition that grounds (all) knowledge in experience – the possibility of innate or spontaneously arising knowledge is either denied outright or strongly downplayed – with an emphasis on the role of sensory perception[11]. Vision being often taken as the dominant of the human senses, so "image" is seen as the most natural metaphor for concepts. Prinz suggests that it is "natural" to interpret John Locke as holding that concepts are mental images, whilst noting that this interpretation is controversial (2004, p. 25). He offers as a surer example John Hume, as well as George Berkeley (with his strong rejection of definitionism, as the quote from him shows); not coincidentally, all three are associated with the tradition of British empiricism.

If definitions are often too restrictive, images face the opposite problem. Images, particularly when taken out of context, are by their nature ambiguous. Both Fodor and Prinz cite Wittgenstein's example of the man climbing a hill, given as a footnote to the section quoted (in part) above:

> I see a picture; it represents an old man walking up a steep path leaning on a stick. – How? Might it not have looked just the same if he had been sliding downhill in that position? Perhaps a Martian would describe the picture so. I do not need to explain why *we* do not explain it so (2001, § 139(b)).

No explanation is needed, Wittgenstein would say, because concepts disambiguate what images on their own cannot. Likewise there is his example of the duck-rabbit (2001, Part II, § xi) (see Figure 2.2).

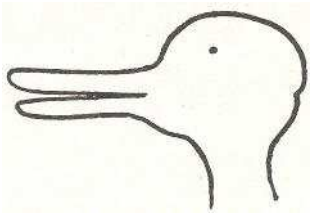

Figure 2.2: Wittgenstein's duck-rabbit: either a duck looking left, or a rabbit looking right.

[11] . . . By which is meant the traditional "five senses": i.e., interoception and proprioception are excluded.

The chief argument against imagism is its traditional reliance on resemblance. Resemblance is well accepted in philosophy of mind to be problematic; the standard arguments why this is so are attributed to Nelson Goodman:

> The plain fact is that a picture, to represent an object, must be a symbol for it, stand for it, refer to it; and that no degree of resemblance is sufficient to establish the requisite relationship of reference. Nor is resemblance *necessary* for reference; almost anything may stand for anything else. A picture that represents – like a passage that describes – an object refers to and, more particularly, *denotes* it. Denotation is the core of representation and is independent of resemblance (1976, p. 5).

Resemblance is often superficial, and appearances can be deceiving, as Gareth Evans notes:

> Whales do not really belong with the fish they superficially resemble, since the similarity of form and behaviour conceals radical differences of structure and function (1982, p. 1).

In philosophical terms, that one thing resembles another is an *explanandum* not an *explanans*: something to be explained, and not itself an explanation. Putting this another way, what Goodman can be said to establish is not that resemblance is unrelated to representation and reference, but that its customary position in the order of explanation is wrong: A does not represent B because A resembles B; rather A resembles B – at least in some circumstances – because A represents B. Putting this another way again, we find resemblance where we look for it[12]. This will be important when the discussion turns to conceptual spaces theory in Chapter Six.

A corollary note: if similarity is problematic ("a resembles b"), so, too, is identity ("a is the same as b" or "a equals b"). When, after all, are two things *enough* the same to *be* the same, and when are they too different? Similarity and identity are like two sides of a coin; if the one cannot be taken for granted, neither can the other.

2.4 Contemporary Discussions: Concepts as (Mental) Representations

> From our assertion that philosophy provides definitions, it must not be inferred that it is the function of the philosopher to compile a dictionary, in the ordinary sense. For the definitions which philosophy is required to provide are of a different kind from those which we expect to find in dictionaries. In a dictionary we look mainly for what may be called explicit definitions; in philosophy, for definitions in use... (Ayer, 2001, p. 48).

If classical definitionism is dead, then Ayer's notion of definitions-in-use remains, at least on Fodor's account and much to his displeasure, very much alive. Ayer goes on in the same passage to explain: "We define a symbol *in use* not by saying it is synonymous with some other symbol, but by showing how the sentences in which it significantly occurs can be translated into equivalent sentences, which

[12] In slogan form: Resemblance does not yield representation; representation yields resemblance.

contain neither the *definiendum* itself, nor any of its synonyms" (2001, p. 49). What he means is something like this: dictionary definitions define a term by providing a synonymous expression; the term can reliably be substituted by the expression and the expression by the term. Definitions-in-use provide synonymous expressions on the fly, ones that are normally implicit until we draw attention to them. Although Fodor does not name names, clearly he has someone like Robert Brandom (2001) in his sights.

There are many modern variations on definitionism that, in one way or another, seek to move definitions beyond an explicit, fixed set of necessary and sufficient conditions. A common move is to say that concepts are definitions (on something like the classical account) "plus _____", where various proposals are made to fill the blank. One particularly interesting proposal from Ray Jackendoff talks of three sorts of conditions: necessary, "graded" (basically, sufficiency conditions along a continuum)[13], and typical (i.e., subject to exceptions) (1985, p. 121). Notice how, in this way, Jackendoff is moving definitionism closer to imagism, with its emphasis on similarity (however that is measured) rather than strict identity. The unified conceptual space theory presented in Chapter Six is, from one perspective at least, itself an updated form of definitionism, as explicitly noted in sections 7.2 and 7.5. It owes more than a passing nod to Jackendoff.

Likewise if few contemporary philosophers would subscribe to a Berkeleyan account[14], then prototype theories of concepts and the closely related *similarity space* theories remain popular. According to contemporary prototype theories, concepts are less images and more bundles of features and probabilities[15], organized around best (prototypical) examples, which may be either actual instances (*exemplars*) or idealized abstractions from actual instances.

Similarity space theories locate concepts within *conceptual spaces* that are the analogue of physical volumes. Rather than length, width, and height, the dimensions are integral properties along which concepts may differ: e.g., in the context of colour, those would be hue, saturation, and brightness. Similarity is then measured in terms of distance within this space. Perhaps the best-known similarity space theorist is Peter Gärdenfors, whose *conceptual spaces theory* of concepts provides the foundation for the unified conceptual space theory.

What all of these accounts share in common with each other and with the classical accounts is a commitment to concepts as abstract objects, specifically representations, often qualified as *mental representations* (in contrast to presumably non-mental representations like a painting or a sculpture – though for considerations why such a distinction is problematic, see sections 1.5 and 2.6). On some accounts, the focus is toward (or exclusively on) *symbols* or *symbolic representations*; on others, *iconic* (image-like) *representations.* But the commitment to representation itself is largely unquestioned: concepts just are the kind of things that stand in for (represent) objects, structures, and other regularities in the world we experience.

[13]"These conditions specify a focal or central value for a continuously variable attribute; the most secure positive judgments are for those examples that lie relatively close to the focal value of the attribute in question. I call such conditions *centrality* conditions" (Jackendoff, 1985, p. 121).

[14]Berkeley (1999), for example, rejects any substantive class/instance distinction altogether, as his triangle example shows: for Berkeley, there is no general concept of a triangle that is not itself a representation or image of a particular triangle. I, too, will want to reject any substantive class/instance distinction (see sections 6.1 and 7.2), but whereas Berkeley makes everything (effectively) an instance, I will attempt to make everything (within limits of practical expansion) a class.

[15]In formal terms, "most concepts... are complex representations whose structure encodes a statistical analysis of the properties their members tend to have" (Laurence and Margolis, 2002).

Rather than try, at this point, to give an exhaustive account of the broad range of contemporary theories, I will look at two representative examples, which are also among the most hotly debated.

2.4.1 Informational Atomism

> ... Content is constituted by some sort of nomic[16], mind-world relation. Correspondingly, having a concept (concept possession) is constituted, at least in part, by *being in* some sort of nomic, mind-world relation. ... Most lexical concepts have no internal structure (Fodor, 1998, p. 121).

If Fodor is keen to reject definitionism in any of its forms, there are aspects of the classical definitionist account that he eagerly shares, both its tendency to associate concepts with words of a language (paradigmatically symbolic) and the central role it offers to compositionality (something that, according to Fodor, other theories of concepts downplay or ignore). Thought is propositionally structured, and "the key to the compositionality of thoughts is that they have concepts as their constituents" (Fodor, 2008, p. 20). Likewise complex (phrasal) concepts are composed out of simple ones. If something is a "brown cow", then it automatically follows that that something *must* be brown: **BROWN COW -> BROWN** as a matter of analytic (as opposed to synthetic) truth. Note that this is independent of Quine's (1951) attack on the analytic/synthetic distinction, since it is meant to be logically true regardless of the concepts involved or any facts about the world – indeed, regardless of whether one possesses any concepts besides **BROWN** and **COW**, since on Fodor's account, it is conceivable that one might not:

> Indeed, it's plausible prima facie that "a" might refer to *a* even if there are *no* other symbols. The whole truth about a language might be that its only well-formed expression is "John" and that "John" refers to *John*. I do think that uncorrupted intuition supports this sort of view; the fact that "John" refers to *John* doesn't *seem* to depend on, as it might be, such facts as that "dog" refers to *dogs* (2008, p. 54).

A useful way to distinguish Fodor's approach from definitionism is that, on the classical definitionist account, all or most concepts compose both *upward* and *downward*: that is, they both join together to form complex concepts and can themselves be decomposed *into* concepts, in the manner that **BACHELOR** is supposed to decompose into **UNMARRIED** and **MAN**. For Fodor, most concepts – roughly, the lexical ones[17] – have no internal structure; they are atomic symbols: hence, the atomism of informational atomism. Therefore, for Fodor, concepts compose upward but *not* downward.

Care should be taken here to read Fodor correctly. He does *not* mean that these atomic symbols are completely unstructured; they must have structure of some kind to be distinguishable one from another. Rather, what structure they have is irrelevant to their role as concepts; they have no internal *conceptual* structure: i.e., their structure is non-conceptual. By analogy, consider the way that the phrase "brown cow" can be broken down into the words "brown" and "cow", but "cow" (unlike "battleship") cannot be broken down into words, only letters; while "cow" can be broken into the letters "c", "o" and "w", but "c" cannot be broken into further letters, only into strokes of a pencil or pen.

[16] Law-like.

[17] There are exceptions: e.g., **BATTLESHIP** which is comprised of **BATTLE** and **SHIP**.

The other half of informational atomism is the ***informational*** *semantics*. The concept **DOG** is a concept *of* dogs not because of any internal structure – in terms of relevant structure, it has none – but because it reliably applies to and only to *dogs* (except, say, on a dark night, or when the agent is drunk, or under any other circumstances that mediate against normal, correct identification). Note that nothing further can be said about the relationship between the concept and its referent: that relationship is the foundation from which everything else is built. Some version of direct realist metaphysics is assumed. (An indirect realist would look for more to be said about the relationship, and an anti-realist would either deny the matching-up altogether – taking a constructivist or enactivist route (see Section 6.1.2) – or at least look for more to be said about the referent.)

2.4.2 Proxytypes: Informational Semantics Without the Atomism

> If concepts are proxytypes, thinking is a simulation process.... Tokening a proxytype is generally tantamount to entering a perceptual state of the kind one would be in if one were to experience the thing it represents. One can simulate the manipulation of real objects by manipulating proxytypes of them in their absence. The term "proxytype" conveys the idea that perceptually derived representations function as proxies in such simulations (Prinz, 2004, p. 150).
>
> ... Proxytypes theory is like informational atomism without the atomism (Prinz, 2004, p. 164).

If Prinz is keen to distance himself from "pictures in the mind", he also holds that "... imagism is less wrong than is often assumed" (2004, p. 103). He shares with imagism its empiricist commitment to grounding all concepts in perception, or, as the Perceptual-Priority Hypothesis expresses it: "nothing is in the intellect that is not first in the senses..." (2004, p. 106). Likewise, despite his commitment to informational semantics, he is committed to a role for similarity in reference. This is because, unlike Jerry Fodor, who sees informational semantics as leading "naturally" to conceptual atomism (1998, p. 156), Prinz denies atomism. Although concepts in the mind attach to their referents in the world because of a law-like relation between them – hence the informational semantics – they do so, at least in part, because of the relationship between the structure of the concepts and the structure of the referents.

According to proxytype theory, thinking in general – and conceptualizing in particular – is fundamentally a process of simulation. Prinz references Lawrence Barsalou (1999), but a recent paper by Victor Gallese and George Lakoff makes the point even more boldly, with the slogans "understanding is imagination" (2005, p. 456) and "imagination is mental simulation" (2005, p. 458).

Although Prinz would be the first to acknowledge that concepts-as-proxytypes are more unlike their referents than they are like them – a concept of a bird has little in common with a bird – nonetheless his proxytypes perform as they do in mental simulations precisely because of certain, albeit superficial, structural similarities they bear to their referents: hence the talk of "scale models". As a bird has wings and feathers, so one's **BIRD** concept has **WINGS** and **FEATHERS** among its other content[18].

[18] The structural isomorphism does not stop there. Just as concepts bear a certain isomorphism to their referents, so likewise the relationship between concepts is isomorphic to the relationship between their referents. If I **THROW** a **STONE** at a **BIRD** and **HIT** it in simulation, the outcome predicted by the simulation is what I would expect to happen were I actually to throw a stone at a bird and hit it, not in simulation.

The name "proxytype" is deliberately evocative not just of proxies but of prototypes. But proxytypes differ from prototypes in several critical ways (Prinz, 2004, p. 156):

- Prototypes do not typically contain linguistic information: e.g., category names. Proxytypes do. Prototypes are iconic representations; proxytypes are a mixture of iconic and symbolic representations.
- Prototype theories do not typically offer a theory of primitive elements from which non-primitive concepts are derived. Proxytype theory holds that concepts are derived from perceptual primitives, or what might be called on some accounts (such as the one offered in chapters Six and Seven) proto-concepts.
- Prototype theories typically are committed to using similarity to account for reference. Although proxytype theory sees similarity as essential to reference, it turns the order of explanation around in the manner described earlier: *because* the concept attaches to a particular referent, *therefore* the concept is made to resemble, in some fashion, its referent. This is more or less the same move I will claim Peter Gärdenfors makes with his conceptual spaces theory (see the introduction to Chapter Six).

Proxytypes revise and extend prototypes[19], just as prototypes sought to revise and modernize imagism. Prinz's main departure from classical imagism and classical empiricism is over their emphasis on concepts as "conscious pictures" (2004, p. 139). Concepts-as-proxytypes are *multi-modal* entities, deriving their content from all of the sensory modalities, not just vision. Likewise, they derive from perceptual representations that are largely unconscious, implying that they themselves may not be, much of the time, consciously introspectible.

2.5 Dissenting Voices: Concepts as Abilities

> Abilities are prior to theories, they say. Competence is prior to content. In particular, *knowing how* is the paradigm cognitive state and it is prior to *knowing that* in the order of intentional explanation. Therefore, don't think of thinking as being *about* the world; think of thinking as being *in* the world. Do not say that the world is what makes your thoughts true (or false); say that the world is what makes your actions succeed or fail. Skepticism obligingly dissolves: Maybe you can't tell whether your beliefs are true, but certainly you can often tell whether your plans succeed. Often enough, it *hurts* when they don't (Fodor, 2008, p. 10).

Besides being a realist, Fodor is a rationalist, in the tradition of Descartes: knowledge in general and conceptual knowledge in particular is ultimately derived not from perception but from reason and logic. He claims general consensus that "the empiricist project failed", even while acknowledging Prinz's work (2008, p. 28). Prinz, for his part, has claimed that even a rationalist like Fodor must, in the end, come around to some form of empiricism (Prinz, 2007). Yet for all their substantive (as opposed to terminological) disagreements, they have areas of substantive overlap. They are both committed to reifying concepts as mental representations, abstract "objects" whose properties (including compositionality) can be set out in much the same way as for physical objects.

[19] . . . As, again, does conceptual spaces theory (see Section 6.1.1).

Concepts are instances of *knowledge that*[20]: so e.g. the concept **DOG** is the knowledge (among, perhaps, many other things) that *that thing* (pointing) is a dog[21]. Conceptual knowledge, on either Fodor's or Prinz's account, is knowledge that, should I engage in such-and-such actions, such-and-such consequences will result. As Fodor says, "You can't think of a plan of action unless you can think about how the world would be if the action were to succeed; and thinking *the world will be such and such if all goes well* is thinking the kind of thing that can be true or false" (2008, p. 13). *Knowledge that* is reflective knowledge: to possess it, one must be aware of having it and aware of what it entails. (If I know that cats are mammals, I must know that there exist such things as cats, and that they relate to mammals in a certain way.)

There is, however, a competing minority tradition, which sees concepts, in the entirety or at least in the first instance, not as *knowledge that* but as *knowledge how* – a position which Ryle himself would surely have been inclined toward. Unlike *knowledge that,* one can possess *knowledge how* without being aware of having it, and in particular without being able to decompose it into component bits of knowledge. (I know how to ride a bicycle, but I cannot tell you how I am able to do so.) One does not need to think about *knowledge how*; one can just get on with using it. To the extent we reflect on *knowledge how*, it becomes *knowledge that.*

On this competing tradition, the concept **DOG** is not the knowledge that *that thing* (pointing) is a dog; the concept **DOG** is the ability to relate to *that thing* as a dog (which includes being able to point it out). Conceptual knowledge is not some reified informational structure but a skillful activity[22]. Advocates generally share a disregard for representations. Concepts do not represent the world, for that would be to imply an illicit or unnecessary distance between concepts and the lived world, a stepping back from rather than immersion in the world.

Fodor takes the advocates of "definitions in use" to task precisely on this point: "Philosophers who accept the idea that having(/learning) a word(/concept) is knowing its definition-in-use just about invariably also assume that the kind of knowing that's pertinent is 'knowing how' rather than 'knowing that'" (2008, p. 35). That said, the view of concepts as abilities has been subscribed to by many who would not describe concepts as definitions at all. As before, the goal here is not in any way to provide an exhaustive review but rather a few representative examples, all of which will feature in important ways in the discussions to follow.

[20] I am drawing here, of course, on Gilbert Ryle's (1949) *knowledge that/knowledge how* distinction.

[21] . . . What is often termed an *ostensive definition*. Note that ostensive definitions are, for all their apparent precision, unavoidably vague, involving as they do a "close enough" match (see (Novak, 2005, p. 345)).

[22] . . . A phrase that will come up again in the discussion of enaction (see Section 6.1.2).

2.5.1 Gottlob Frege

> The word "concept" is used in various ways; its sense is sometimes psychological, sometimes logical, and sometimes perhaps a confused mixture of both. Since this license exists, it is natural to restrict it by requiring that when once a usage is adopted it should be maintained. What I decided was to keep strictly to a purely logical use... (Frege, 1951, p. 168).
>
> In the enquiry that follows, I have kept to three fundamental principles: always to separate sharply the psychological from the logical, the subjective from the objective; never to ask for the meaning of a word in isolation, but only in the context of a proposition; never to lose sight of the distinction between concept and object (Frege, 1980, p. x).

Figure 2.3: Gottlob Frege. (Photo downloaded from Wikimedia Commons: http://commons.wikimedia.org/.)

Although Frege (see Figure 2.3) may not be the first name that comes to mind when talking about concepts as abilities – Gareth Evans (1982), Christopher Peacocke (1992), or the aforementioned Robert Brandom (2001) would be more typical citations – nonetheless he does seem naturally read in this light, making him their intellectual forerunner. For Frege, concepts have no meaning outside of their specific application. They are intimately bound up with language. They are *logical simples*, beyond any but the most minimal explanation; their nature, other than their ability to predicate, can at best be hinted at. ("Hint" – *Wink* in German – is, as Kelly Dean Jolley notes (2007), a word that Frege uses a lot.) They are known by their use. This makes them even more bare than Fodor's atoms, for Fodor's atoms are still symbols and still mental representations, of which much can be said.

It should be allowed that Frege is intending "concept" (*Begriff)* and "object" (*Gegenstand*) in a technical sense, as the above quote suggests – one that permits of other appropriate uses of these words (though he does think he has "got hold of a distinction of the highest importance" (1951, p. 179)). One could argue that his concept of a concept is incommensurable with the one I am describing in this work. Nonetheless, Frege is widely cited in the literature on concepts, and so it seems reasonable to suggest that his target is the same as mine; Frege's own admonition on this point bears reproducing here: "... we cannot come to an understanding with one another apart from language, and so in the end we must always rely on other people's understanding words, inflexions, and sentence-construction in essentially the same way as ourselves" (1951, p. 171).

Nonetheless, Frege's usage requires unpacking. For Frege, concepts arise in the context of propositions, where propositions are thoughts that can be expressed as truth-functional statements of a language[23]. A proposition consists of an *object expression* and a *concept expression*. Just as a subject and a predicate together make a sentence, so an object expression and a concept expression together make a proposition. Object expressions ("my dog Fella", "the planet Venus") are *saturated* (complete); concept expressions ("is now dead", "is the same as the Evening Star") are *unsaturated* (incomplete, requiring the addition of an object). Object expressions identify objects; concept expressions identify concepts. Objects are predicated about; concepts do the predicating. Things are far from straightforward however because, as Frege acknowledges, first-order (or *first-level*) concepts predicate about objects in a similar way to how second-order (or *second-level*) concepts – roughly, concepts of concepts – relate to first-order ones.

Three points are critical:

- Objects, for Frege, can never be concepts, nor concepts objects. They are fundamentally different things (as are first- and second-order concepts). As Kelly Dean Jolley writes, "By calling objects *saturated* and concepts *unsaturated*, Frege seems to suggest that they are two species of one genus... [but] the distinction between *saturated* and *unsaturated* is... *a distinction without a genus"* (2007, p. 68).

- Specific instances as identified by proper names ("my dog Fella") or definite descriptions ("the queen of England") can never be treated as classes *contra* the discussion in Section 2.1, for the former are objects while the latter fall under concepts.

- Object expressions and concept expressions do not come together to form propositions; rather, propositions come first, and can be broken down into object expressions and concept expressions. Propositions are primary; object expressions and concept expressions are secondary. One of the consequences of this is Frege's *context principle*, quoted above: one should never ask for the meaning of a word except in the context of a specific proposition in which it is being used.

In consequence of these, one cannot identify a concept with an expression such as "the concept **DOG**" (or "the concept **CONCEPT**", for that matter). To allow otherwise would be to allow such propositions as "the concept **DOG** is a concept easily attained" (with apologies to Benno Kerry), and that would be to make an object out of a concept, which one cannot do. To be sure, "the

[23] By some accounts (e.g., (Torey, 2009)), thoughts "just are" structured out of words, and so these propositions are necessarily linguistically structured: not in a Fodor-type language of thought but in words of actual spoken language. It is not clear to me that Frege is making this commitment; and in any case, I will not. Henceforth where I use the term "proposition", no such linguistic structure should be implied.

concept **DOG**" in that statement picks out *something*, but what it picks out is not a concept but only an object masquerading as a concept. Because concepts for Frege are tied so tightly to their usage, and because they are clearly not representations, nor any sort of objects (in the technical or non-technical sense), it seems appropriate to think of them as abilities.

2.5.2 Gareth Evans

> It seems to me that there must be a sense in which thoughts are structured. The thought that John is happy has something in common with the thought that Harry is happy, and the thought that John is happy has something in common with the thought that John is sad. This might seem to lead immediately to the idea of a language of thought, and it may be that some of the proponents of that idea intend no more by it than I do here. However, I certainly do not wish to be committed to the idea that having thoughts involves the subject's using, manipulating, or apprehending *symbols*.... I should prefer to explain the sense in which thoughts are structured, not in terms of their being composed of several distinct *elements*, but in terms of their being a complex of the exercise of several distinct conceptual *abilities*. Thus someone who thinks that John is happy and that Harry is happy exercises on two occasions the conceptual ability which we call "possessing the concept of happiness" (Evans, 1982, p. 100).

Although Evans is, perhaps, more than most philosophers, associated with the idea of concepts-as-abilities, his *magnus opus* (and only book, edited and published posthumously) is concerned in the main not with concepts as such but with *singular terms*[24], which include proper names and, by Frege's account at least, definite descriptions. Because Evans is keen to cast himself by and large within the Fregean tradition[25], singular terms and concepts are going to be quite separate things. Furthermore, Evans specifically sets out in his introduction to ignore questions of ontology: his interest is less in what things are than in what can coherently be said about them in the context of how they are used.

Nonetheless, though Evans does not want to commit to structured (i.e., conceptual) thought *not* being symbolic or representational, he clearly is more comfortable talking about concepts as abilities. His *Generality Constraint* – "if a subject can be credited with the thought that a is F, then he must have the conceptual resources for entertaining the thought that a is G, for every property of being G of which he has a conception" (1982, p. 104) – will be discussed at more length in Section 3.1, where it will be taken to ascribe certain essential, untendentious (arguably non-contestable) properties to concepts. But properly speaking, for Evans these are properties not of concepts *per se* but of conceptual abilities. Conceptual abilities abstract away from particular instances in order to then be applied back to particular instances:

> There must be a capacity which, when combined with a knowledge of what it is in general for an object to be F, yields the ability to entertain the thought that a is F, or at least a knowledge of what it is, or would be, for a to be F (Evans, 1982, p. 103).

So concepts, whatever exactly they are, are not specific to a particular context; yet their application is always specific to a particular context, which is possible *precisely because* they are not specific

[24] Evans also talks about *referring expressions*, by which he means the same thing.

[25] He is echoing Frege quite deliberately when he says e.g. that "...the understanding of a word is manifested only in the understanding of sentences..." (1982, p. 102).

to a context. A **DOG** concept that applied only to a particular dog on a particular occasion – if there could be such a thing – would be of no usefulness at all.

2.5.3 Alva Noë

> If the animal is present *in* the world, with access to environmental detail by movements... then why does it need to go to the trouble of producing internal representations good enough to enable it, so to speak, to act as if the world were not immediately present? Surely we sometimes need to think about the world in the world's absence (when it's dark, say, or when we're blind, or not at the location we're interested in), and for such purposes we must (in some sense) represent the world in thought. But what reason is there to think that this is the case in standard perceptual contexts? In many situations, we need only move our eyes, or move our head, or turn around, to get whatever information we need about the environment. How many bookshelves are there in your room? You don't need to have an internal representation to answer; you need only to be able to turn around and take a look (Noë, 2004, p. 22).

Like Prinz, Alva Noë is trying to offer an updated empiricism (though he does not specifically identify himself that way). In Noë's case, it is an empiricism that emphasizes *active* perception: in particular, the link between motor actions and sensory consequences, if I do *this* then *that* will happen. Sets of such expectations he calls *sensorimotor profiles.*

Although much of their empiricism overlaps, there is a difference of emphasis: whereas for Prinz much of the focus is on off-line simulation, for Noë it is on immediate engagement. Noë draws inspiration from J.J. Gibson's work on affordances (1986), noting that according to Gibson, "for the active animal, the ground is directly perceived as walk-uponable, and the tree stump as sit-uponable" (2004, p. 21). That is, the ground directly *affords* walking and the tree stump sitting.

For Prinz, concepts are mental representations; for Noë, concepts are, like the perceptions they derive from, skillful bodily activities. The key knowledge is not *knowing that* but *knowing how* (see e.g. (2004, p. 11)). Noë's insights about grounding concepts in sensorimotor activity will feature prominently in Chapter Seven.

As with Frege and his "concepts" and "objects", Noë is using "perception" in a technical sense that emphasizes active experience: truly passive perception, for Noë, is a contradiction. An incautious reading might miss this. Bacteria, insects, and other simple organisms do not perceive anything, at least insofar as we cannot attribute them some minimal level of active understanding:

> Blind creatures may be capable of thought, but thoughtless creatures could never be capable of sight, or of any genuine content-bearing perceptual experience. Perception and perceptual consciousness are types of thoughtful, knowledgeable activity (Noë, 2004, p. 3).

Again:

> ...Merely to be given visual impressions is not yet to be made to see. To see one must have visual impressions that one *understands* (Noë, 2004, p. 5).

Noë does not necessarily mean that agents must have *conceptual* understanding in order to perceive. On this issue, Noë draws a pragmatic line:

> The understanding of concepts is usually supposed to be a paradigm of personal-level accomplishment. But just as there is no sharp line between the personal and the subpersonal, so there may be no sharp line between the conceptual and the nonconceptual. Indeed, it may be that sensorimotor skills deserve to be thought of as primitive conceptual skills, even if, as is frequently the case, they are subpersonal (Noë, 2004, p. 31).

Noë states all of this in oddly qualified terms. Yet it is clear from elsewhere in his book that this is not a position he is raising as a mere possibility but one he actively endorses. Note one consequence of Noë's approach to concepts: they are no longer tied directly to language. *Contra* Frege, language and concepts pull apart. This is an idea I will return to several times, particularly in Section 3.3.2.

2.6 Symbols and Other Representations

I started this chapter by considering the folk understanding of concepts as structured ideas, ones that abstract away from particular instances or experiences. Philosophy, in trying to make this more precise, has historically tended to dress concepts in representational language, so that concepts stand or *stand in* for aspects of the world: either as symbolic representations as e.g. in the case of classical definitionism or iconic representations as in e.g. the case of classical imagism. Meanwhile, a longstanding minority opinion rejects a representational approach as a misguided stepping back from the world; agents are always *in* the world, directly engaged with it, and their concepts should be understood likewise: not things to be described so much as hinted at, not units of information to assemble/disassemble but skills to be acquired or refined.

The debate between representationalists and anti-representationalists shows no signs of abating. One might think that, given its intensity, the ground rules for the debate would be well established. Yet both "symbols" and "representations" are terms often used by both sides without being defined, as if their definitions were clear and consistent, and universally agreed upon. They are not. Even where people may think they are being clear what they mean, their usage can be problematic. Of course nobody owns the "correct" definition of these terms; but still, a careful definition of terms seems like a good place to begin. Therefore, in order to defend the position I myself want to take, I will need to first make a digression.

2.6.1 Symbols (Symbolic Representations)

> A symbol in a model is arbitrary if there is no obvious relation between the mark or sound we use to designate that symbol and the things represented by implementations of that symbol (or realizers of it, or objects onto which that symbol is mapped during modeling, etc.) (Rupert, 2009a, p. 221).

For the discussion that follows, I will take "symbol" and "symbolic representation" to mean the same thing.

In lay terms, a symbol is something that by virtue of convention stands in for something else. The "something" is normally relatively simple, the "something else" relatively complex. It can be anything readily re-identifiable (and ideally reproducible), whose simplicity of form belies the rich set of meanings that have become attached to it. A symbol can be a mark on a page ¶ (symbol for new paragraph), a clenched fist stenciled on a wall (symbol for resistance), a wooden cross (symbol to Christians for resurrection). Regardless of how it is instantiated, the symbol should normally be visually discernible or at least visually expressible: something that can be expressed only as sound, smell, taste, or tactile sensation would not qualify as a symbol.

Once again, philosophers use "symbol" in a related but slightly different and more formal way, taking their cue from mathematicians, computer scientists, and so on. So philosophers typically describe symbols as[26]:

1. Amodal: independent of and not grounded in any sensory *modality*, be it sight, sound, taste, smell, or touch.

2. Interpretable fully independently of context, with a strict separation between syntax (form) and semantics (meaning). In linguistic terms, they are *context-free* not *context-sensitive.*

3. Discrete (individuable), not continuous.

4. Fully arbitrary (see the Robert Rupert quote above), so that any symbol can in principle be substituted for any other symbol, providing only that the usage is consistent.

5. Observer independent. That a symbol is a symbol is intrinsic to its nature and not dependent on any agent to assign it its identity or meaning[27].

A familiar image is that of a computer applying purely syntactic rules to strings of "meaningless" symbols to generate new strings of symbols, in the manner of Typographical Number Theory (Hofstadter, 2000, p. 204 ff.) or John Searle's (1980) famous Chinese Room thought experiment. The computer, it is said, does not understand the symbols it is operating over – any more than, for advocates of e.g. the Physical Symbol System hypothesis (Newell, 1980), does the brain's "machinery" (i.e., neurons and synapses). In what sense, after all, *could* the computer be said to understand what it is doing? The obvious answer for many is, none.

As Mike Wheeler notes, much of classical cognitive science took representations "just to be" symbols, in this naïve sense:

> In classical cognitive science a representation is either an atomic symbol or a complex molecular structure constructed through the systematic recombination of atomic symbols according to syntactic rules. The content of a molecular representation is a function of the contents of the constituent symbols plus the syntactic structure of the complex formula. In other words... classical representations, in a manner familiar from natural and artificial languages, feature a combinatorial syntax and semantics (Wheeler, 2005, p. 62).

[26] A longer but related list can be found in (Harnad, 1990d). Note that Harnad's list should probably be seen as more demanding than this one. Therefore if this list raises difficulties for symbols, so does his. Likewise compare this list with the nearly identical properties given by (Gallese and Lakoff, 2005, p. 455) for concepts as described on "first-generation cognitive science" accounts "influenced by the analytic tradition of philosophy of language".

[27] This list relates to the corresponding lists in sections 2.6.1.1 and 2.6.1.2 as follows: the second list shows point by point what is problematic about each of these claims, while the third list shows point by point how each of these claims *should* be understood.

2.6.1.1 Symbol Problems

At the same time, all of the above properties are known to be problematic, to the extent that one might well wonder if most people – even philosophers – really mean by symbols what they think they mean.

1. Symbols have to be grounded *somehow*, if not modally then in some other fashion: logically, semantics is never free floating. (In a slogan: meaning does not come for free.) This is, in effect, Stevan Harnad's (1990d) symbol grounding problem.

2. Symbols are only ever meaningful with respect to some context. At minimum there is always a shared social context in which they are learned and applied (see (Harvey, 1992) and below). Remove them too far from their context of origin, and they cease to function as symbols, reduced to meaningless markings or the like. Furthermore, the separability of syntax from semantics is far from untendentious: e.g., Gärdenfors writes, "Semantics is primary to syntax and partly determines it (syntax cannot be described independently of semantics)", a position that, he notes, "... is anathema to the Chomskian tradition within linguistics" (2004, p. 165) – not to mention the cognitivist and rationalist traditions in cognitive science.

3. Symbols evolve: they change over time[28]. At what point does a symbol cease being one symbol and become another? When did the Sanskrit good-luck symbol, the *svastika,* become the German swastika of racial purity – or do they remain distinct symbols? Furthermore, symbols never stand on their own but exist in relation to other symbols, some of which may be quite similar. At what point are two symbols not two different symbols but the same one? The answer is unclear and seems to depend upon the context of application.

   Consider these different but obviously related ways of writing the number "one" (see Figure 2.4). Likewise consider the two ways of expressing the number two in German: *zwei* or *zwo*; or compare the German *zwo* with the Swedish *två* which look quite different but are pronounced and mean the same. Are they the same symbol with slightly different expressions or different symbols that happen to be closely related?

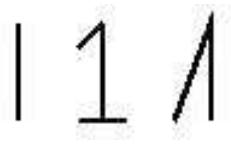

Figure 2.4: Three ways of writing the number "one".

4. The relationship between the form (syntax) and the meaning (semantics) of a symbol often appears to be *non*-arbitrary, which, again, tells against their separability. So for example, the symbol "1" is reminiscent of the marks used in pre-ordinal counting, from which it very probably derived (see Figure 2.5), and the symbol "3" has three prongs. The clenched-fist symbol resembles a clenched fist, and the Christian cross is meant to resemble a wooden cross. In Chinese, as Harnad (1990d) notes, the line between characters as letters, characters as words, and characters as pictures is greatly blurred. In any case, the form a symbol takes always comes with a history.

[28] Compare the discussion of conceptual change in Section 3.2.4.

Figure 2.5: Pre-ordinal counting.

5. It is, to say the least, far from clear what it means for a symbol to be a symbol independent of an agent to interpret and employ it as such. Inman Harvey, among others, has long argued that it is in the nature of a symbol, as a form of representation, that it is not a symbol in the absence of an agent to give it meaning.

   Failure to acknowledge the role of the observer in the act of representing leads to confusion. And yet, "the underlying assumption of many is that a real world exists independently of any observer; and that symbols are entities that can 'stand for' objects in this real world in some abstract and absolute sense. In practice, the role of the observer in the act of representing something is ignored" (1992, p. 5). His admonition on the sloppy use of "representation" bears frequent repeating: "The gun I reach for when I hear the word *representation* has this engraved on it: 'When P is used by Q to represent R to S, *who is* Q *and who is* S*?*"(1992, p. 7)

To return to the computer example: if Inman's observation is correct – and I think it is – this is why a computer is not, on its own, and contrary to the usual interpretation, doing symbol processing (a point that Terry Winograd and C.F. Flores elaborate in (Winograd and Flores, 1986)). The human agent is necessary to give meaning to the signs being manipulated by the computer, and without that agent as part of the process, those signs never become symbols, even in the impoverished sense of number crunching. We treat computers as idealized machines, operating on their own, neither embedded in an environment nor embodied in any particular form (when in fact they are both), unable to make a mistake (which they can, and do), for the same reason we treat symbols as amodal/discrete/context-free/arbitrary – likewise idealizations – because we find it conceptually useful, even, perhaps, as I shall consider in Section 5.2.3, conceptually necessary.

2.6.1.2 Symbol Solutions

A lot of the problems with symbols go away once one rejects the principle of observer independence and takes the remaining four properties as idealizations rather than absolutes: to wit, a symbol is (or is recognizable as) a symbol *to the extent* it meets those requirements. This invites not a hard-and-fast distinction between the symbolic and the sub-symbolic or non-symbolic but something more like a continuum. With all of this in mind, the properties of symbols can be restated as follows:

1. Modally grounded, but in such a way that the links back to the modal grounding may, in practice, be difficult or impossible to reconstruct[29].

2. Consisting of a sign (syntax) – meaning (semantics) dyad (*per* Wittgenstein) where, to some practical extent, the sign can be distinguished from the semantics and the symbol from any particular context of interpretation, without in either case assuming any hard-and-fast

[29] Note how this mirrors the arguments of the concept empiricists, e.g. (Prinz, 2004), for grounding abstract concepts.

boundary.

As Wittgenstein commented, much confusion can arise from confusing the symbol with the sign. "The sign is the part of the symbol perceptible by the senses. Two different symbols can therefore have the sign... in common – they then signify in different ways.... In the language of everyday life it very often happens that the same word signifies in two different ways and therefore belongs to two different symbols or that two words, which signify in different ways, are apparently applied in the same way in the proposition" (Wittgenstein, 1922, § 3.323).

3. Individuable from other symbols or from a non-symbolic background, with the caveat that the individuability masks an underlying continuity: again, a pragmatic boundary, not a hard-and-fast one. That boundary may shift depending on context of application and over time (see e.g. (Harnad, 1990a)).

4. Possessing an *apparent* arbitrariness of form, precisely in relation to the extent to which the symbol has abstracted away from any particular context of interpretation. Rather than form being arbitrary in any ultimate sense, rather the historical relation between form and meaning has been lost or obscured.

5. Observer-dependent: requiring someone for whom the symbol is symbolizing, and someone the symbol is symbolizing to. (These could be one and the same.) Lacking this, things are not "symbol-like" or "sub-symbolic"; they are not symbols at all.

2.6.2 Iconic Representations

Contrast the image of a cat (such as a photograph) with a real cat. The real cat is a mammal, furry, alive, eighteen inches long (say), and composed of flesh and blood, while the image is not a mammal, not furry, not alive, five inches long (say), and composed of paper and Kodak chemicals (Goldberg and Pessin, 1977, p. 74).

If symbolic representations make binary distinctions – something either satisfies a definition or it does not – iconic representations make analogue (graded) distinctions. The standard metaphors for iconic representations are "pictures in the mind" or Prinz's scale models. Pictures differ from symbols in that pictures are neither meaningless nor arbitrary; to recognize something as a picture – even a very abstract example of modern art – is to assign it some minimal interpretation, which is another way of saying that form (syntax) and meaning (semantics) go hand in hand. One consequence is that iconic representations typically lack the combinatorial syntax and semantics that Wheeler discusses.

In common parlance, representation is all about likeness: a painting of a dog should look like a dog. That is to say, the folk notion of representation is closest to the philosophical notion of iconic representation and in contrast to the usage of "representation" in classical cognitive science. A painting of a dog will make a poor representation of a waterfall unless and until the viewer of the painting can establish the (presumably hidden) likeness between the two: in philosophical terms, a *structural isomorphism* between aspects of the painting and aspects of the waterfall.

Of course, even when the painting is taken to be a painting of a dog, the painting and the dog will not have that much in common, as Sanford Goldberg and Andrew Pessin observe. Among other

things, dogs fetch bones; paintings of dogs do not. So the point is not that the likeness will not break down very quickly – it will – but that, for all that it *is* superficial, it is still sufficient to establish and maintain the connection. Details deemed irrelevant in the representational context are simply ignored: e.g., the nature and arrangement of the dog's internal organs. The lesson to be learned is that we tend to find likeness where we look for it, and having found it, we tend not to look deeper. To repeat the earlier point: resemblance rarely, if ever, explains anything; rather, it indicates something requiring explanation.

2.6.3 Representations on a Continuum

Like their symbolic counterparts, iconic representations are things that stand in place of other things. I propose that iconic representations be understood as symbolic representations with the requirement for (relative) arbitrariness between form and meaning relaxed. (Contrast this with Fodor's account of discursive [symbolic] and iconic representations, according to which discursive representations can *never* be iconic representations and *vice versa* (2008, p. 171).) With symbolic representations, the relationship between form and meaning has, for most practical purposes, been lost; with iconic representations, that relationship is still, to greater or lesser extent, apparent. Putting this another way, symbolic representations can be viewed as an impoverished form of iconic representations: pictures reduced in richness and dimension until they are nothing more than "mere" symbols [30].

What this implies, *pace* Fodor, is that iconic and symbolic representations are different regions along a continuum from the fully non-representational (among other things, no identifiable observer) – what some would call the sub-symbolic – to the fully rarefied: symbols so remote from context that they have lost their ability to function as symbols. Iconic representations are a little more toward the "lower" end of the continuum, symbolic representations more toward the "higher" end. Most of the continuum is, ignoring for the moment the conceptual imposition of the observer, not identifiably representational at all.

What makes a symbol an effective symbol – what makes it recognizable as a symbol in the first place – is the extent to which it abstracts away from any particular context of interpretation, to be applicable across the broadest possible range of contexts. The further abstracted away the symbol is from the initial context(s), the less obvious its relationship back to the initial context(s) will be and the more arbitrary the relationship between form and meaning will appear. What makes an iconic representation an effective representation is *both* the extent to which it abstracts away from any particular context of interpretation *and* the extent to which it retains its links back to particular contexts. Symbols are unstructured relative to the domain of interpretation. Iconic representations are not.

Consider the W.A.S.T.E. symbol from Thomas Pynchon's novel ***The Crying of Lot 49*** (see Figure 2.6). Is it in fact symbolic, or is it iconic – or is it something of both? It is recognizable as a muted horn, but only just. Still, it is more richly structured than many if not most symbols, its form anything *but* arbitrary. At the same time, its relatively bare simplicity aids reproducibility, and so it is graffitied onto the walls of toilets, doodled in notebooks, left within the dust jackets of library books, and so on: a trail of hidden code, for those who know to recognize it. Its very

[30] Compare this to the discussion of conceptual abstraction in Section 7.2.3.

simplicity belies its complexity, as new and deeper meanings are revealed throughout the course of the book.

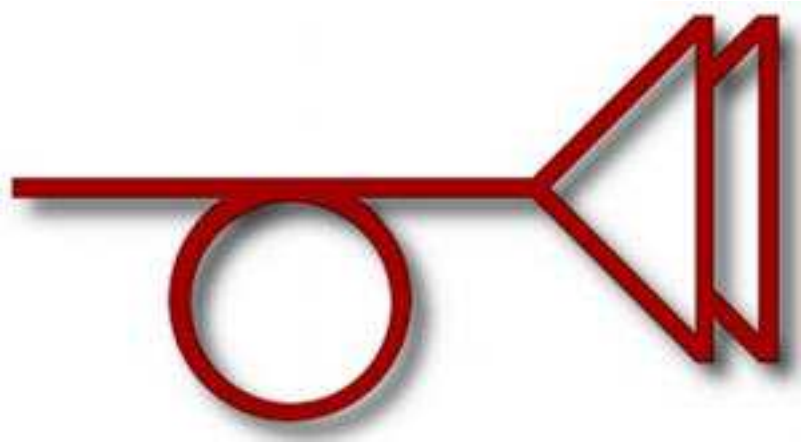

Figure 2.6: The muted-horn symbol, featured in ***The Crying of Lot 49.*** (Photo downloaded from Wikimedia Commons: http://commons.wikimedia.org/.)

2.6.4 Whither Mental Representations?

There is nothing in the definitions I have offered to distinguish mental representations (what some, e.g. (Rupert, 2009a), would call "internal representations") from any other kind of representation (a picture, a painting, a poem, or whatever: things that are commonly called "external representations"). This follows from my reading of Harvey, with his emphasis on the ineliminable role of the observer in the act of representing[31]. A representation, as I intend the term, is not so much a thing as an intentional perspective that an observer takes toward a thing: be it a picture, a painting, a poem – or a belief, a conviction, an idea. Whatever the nature of the target, the relation is the same. Bottom line: representations are relational.

Of course one can meaningfully distinguish mental representations from other representations; I do not wish to deny that. What I am denying is that:

1. Mental representations are ontologically distinct from other representations.
2. Mental representations are (at least without further examination) either historically primary or historically secondary to other representations (i.e., either they come first, or the other ones come first – the latter being Harvey's position).

Note that, on the account I am offering, representations and representeds can never be the same (or never precisely the same): something can never represent itself, in any but a metaphorical sense, because then the relation that defines the representation drops out, just as it does where there is no observer. Of course, one could define terms so that representation and represented *could* be the same; but then the burden is on that person to offer a consistent account of how that is to be done without e.g. inviting an infinite regress (if X represents itself, then it represents itself representing itself, represents itself representing itself representing itself, and so on), or explain why that regress is not a problem.

[31] I should stress that Harvey *is* happy to talk about mental representations, just not internal ones.

Anti-representationalists often talk, if loosely, as if representations should be excised from all discussions on concepts and cognition[32]. I have a more modest proposal: to excise discussion of mental representations or internal representations as offering a vacuous distinction at best, at worst setting up not just a deeply misleading but a false dichotomy, one that invites people to think of inner and outer worlds, of inner experience and outer environment. As I argued in Section 1.5, certain metaphysical starting points encourage people to expropriate the notion of boundary between "inner" and "outer" from the realm of physical volumes – where it has its primary meaning, and its usage is more or less straightforward – to the realm of the cognitive, where its appropriateness is far from clear.

Therefore, for the remainder of this thesis, I will speak of mental representations or internal representations only where others do so, and with no implied endorsement of those terms. I will not speak of external representations at all.

2.7 Between *Knowing How* and *Knowing That*

With these preliminaries out of the way, I am now able to present the first of the central theses of this work and offer a preliminary argument for it, one which shall be developed over the following chapters:

> Despite the seeming unresolvability of the two positions, it is both the case that concepts must be understood as representations (albeit with no "mental" qualifier) and that, logically, concepts must ultimately be something other than representations. In the latter case, concepts will best be understood as non-representational abilities.

In slogan form: the conceptual agent is eternally suspended between (unreflective) *knowledge how* and (reflective) *knowledge that.* Likewise concepts are neither one thing nor the other, a lesson we can take from Gärdenfors' conceptual spaces theory. Push them the one direction, toward explaining "high-level" cognition, and they appear more as representations; push them in the other, toward explaining "low-level" cognition, and they appear more as abilities. Just as iconic and symbolic representations can logically be different positions along a continuum, so, too, can *knowledge how* and *knowledge that.*

The critical distinction to be made – one which is not brought out well if at all in the literature – is between concepts as we reflect upon them *as* concepts, and concepts as we possess and employ them non-reflectively (which would seem to be most of the time). The distinction will receive its fullest expression in Section 5.4, where many if not most contemporary debates in theories of concepts will be seen to line up on one side of the distinction or the other. For now, it is enough to see that, when concepts are the target of our reflections – as they must inevitably be in discussing a theory of concepts! – then representations (as I have defined the term) is both how they present themselves and what they will be. When we are engaged with concepts non-reflectively, then, there being no observer, they cannot be representations. To allow the observer back in would be to commit the so-called *homuncular fallacy*: to assume a "little man" in the mind who is watching everything that is going on, and perhaps another "little man" watching him, and so on, in an unending *homuncular regress.* Such a regress is, for many anti-representationalists, one of the fatal

[32] I have not seen this put so baldly anywhere in writing, but I have often heard it from people face to face.

shortcomings of a representational theory of mind (RTM) – what Fodor has (tendentiously!) called "the only game in town" (Fodor, 2008, p. 113)[33].

2.8 The Ever-Present Observer

I have offered the distinction as though it were straightforward; but the trick is that, the moment we consider what concepts must be when we engage with them non-reflectively, *the very act of reflection* makes them representations. Reflecting on our non-reflecting creates a paradox from which we cannot escape, for to do so would require us to catch ourselves, as it were, in the act. The moment we reflect, the observer intrudes. This brings me to the second of the central theses of this work:

> In these discussions, the observer is always present, if only in the background, and cannot be eliminated.

One may allow that this is true in the case of concepts. But if concepts are, as I suggested in the last chapter, the means by which we structure our experience, then it is far from clear whether they can be bracketed off from everything else. If they cannot be, then the anti-realist and the constructivist have won the argument, and there is no useful sense in which we can step outside our role as observers or describe an observer-independent reality, even if we must, as the anti-realist and the constructivist allow, acknowledge that it exists and continuously constrains our experience.

At the same time, we need means to talk about contexts in which there logically is no observer. If *every* application of a concept required a reflection on the concept as a concept, then the reflection would itself require reflection, and conceptual abilities would never get off the ground.

The situation here is reminiscent of the regress in Lewis Carroll's "What the Tortoise Said to Achilles" (Carroll, 1995). Even Fodor concedes this: "... The 'knowing-how' account of concept possession must be right at least some of the time.... Not all of a mind's transitions from premises to conclusions can be mediated by the application of rules; some of them must be 'immediate', on pain of regress. If, in particular, following a rule itself required reasoning about which rule to follow, or how to follow it, the process of inference could never get started" (2008, p. 36). Fodor's target here is symbolic representations, but the same argument applies equally well to any sort of representation.

On the one hand, it is vital in these discussions to keep the role of the observer in mind and not be fooled into believing that the observer has been eliminated. On the other, there will be plenty of occasions where it is useful to suspend consideration of the observer, and where appropriate I will do so.

The point to hold onto for now is that, when we allow the observer into the foreground, something of even the simplest, pre-conceptual aspects of cognition will appear representational (to an observer!), and any story that does not leave room for representations will be incomplete. When we push the observer into the background, something of even the most complex, abstractly conceptual aspects of cognition will appear not as representations but as abilities. If one asks, but which account is primary? – the answer will depend on where the observer is standing: background or foreground.

[33] It is worth commenting that some researchers – notably Owen Holland (2007) – explicitly endorse a homuncular model, albeit one presumably intended to avoid the regress.

2.9 Toward an Ontology of Theories of Concepts

As said in the introductory chapter, a theory of concepts is any philosophical attempt to approach the question "what is a concept?" within a well-structured, semi-formal framework; to relate concepts to propositional attitudes, most particularly beliefs; put simply, to give an account to the structured nature of our thought: i.e., how our thoughts come to be structured in the ways that first-person introspection and second- and third-person conspecific observation suggest them to be. To the extent that account is conceptually structured – and it is difficult to see how it could not be, at least in substantial part – the question "what is a concept?" becomes "what is our concept of what a concept is?", or more simply "what is our concept of concept?", or just "what is our concept **CONCEPT**?" In this way a theory of concepts becomes a meta-conceptual structure, itself a kind of meta-concept. Discussion of concepts in general becomes inextricably interwoven with discussion of one concept in particular: the most general, overarching concept conceivable.

Theorizing about concepts – i.e., conceptualizing about concepts – is intrinsically a reflective activity. The focus of concepts stops being things in the world – mainly non-concepts – and becomes the concepts themselves. In the arguments of Section 2.7, that makes theorizing about concepts a deeply *representational* activity (and one where one should take care not to confuse the representations with the represenetds!).

2.10 Conclusions

According to folk understandings, concepts are structured ideas that abstract away from particular occasions or instances. Philosophers have attempted to take those folk understandings and offer a more precise account – even while some philosophers have questioned whether a more precise account is possible.

Historically philosophers have tended to think of concepts either as dictionary-like definitions or as abstract images. Definitionism has given way over time to other forms of symbolic representation, imagism to other forms of iconic representation. Nonetheless, some representational theory of concepts is frequently assumed. In recent years, there has been much heated debate whether concepts should best be understood as representations at all or as non-representational abilities.

Representations – particularly mental or internal representations – are frequently discussed without being defined. I have offered a clear definition for both iconic and symbolic representations, while suggesting that such expressions as "mental representations" and "internal representations" are vacuous and best discarded. An essential part of being a representation, *per* Harvey, is an agent to recognize the representation as a representation and actively use it to represent something to someone.

Representational language, of the appropriate kind, is needed for understanding concepts as we reflect upon them, and for putting forward any theory of concepts. On the other hand, one should not confuse the representation of a concept with the concept itself, which must, as I have defined terms, logically, ultimately be something non-representational.

As self-reflective agents, we cannot stop being observers, at least when it comes to concepts. Nonetheless, whether concepts are best understood as representations or abilities depends on where we position the observer.

A number of different theories and names have been discussed. Although the theory put forward starting in Chapter Six will owe the greatest debt to Gärdenfors' conceptual spaces theory, less to Prinz's proxytypes theory, and less again to Fodor's informational atomism, nevertheless it will bear the recognizable mark of all three, and all three names will feature prominently in the discussions to follow. Frege, Evans, and Noë will likewise make their appearances.

Chapter 3

Conceptual Properties

The previous chapter talked about the sorts of things concepts are and talked about their properties or their manner of use only insofar as they were relevant to that basic question. So we saw how they are abstract in their nature, concrete in their application: derived from specific instances or contexts (and hence removed from them) in order to apply back to specific instances or contexts. We saw how they are hierarchical, and how they resemble categories (at the same time that they seem to bridge the type-token divide). We saw how they share the properties of (symbolic or iconic) representations, but only when we, as conceptual agents, are taking a representational stance towards them; and how those representational properties are only idealizations: functionally amodal, relatively individuable, relatively arbitrary, but always relative to the point of view of an observer.

It is time now to look more closely at the properties of concepts, in particular those intrinsic properties they can be expected to have regardless of whether we are reflecting on them or just getting on with using them. Ideally such properties should be, if properly formulated, individually necessary and collectively sufficient. But that, of course, would be to offer a definition of the conceptual; and Section 2.3.1 offered reasons to think that such a definition – a stable and untendentious one, at least – is unlikely to be forthcoming. Furthermore, if the proposal in Section 3.2.4 is correct, that concepts are subject to change, then that would surely apply to our conceptualization of concepts as well. (To what extent did Locke mean by "concept" what a modern philosopher means by "concept"?)

Per Jackendoff's proposal, there may be other, extrinsic properties, commonly associated but not, in a pinch at least, necessary. To understand the difference between intrinsic and extrinsic properties, consider a car. A space in which to sit is an intrinsic property of a car. An artifact designed without any such space would not be a car. A fuel source, or the provision for one at least, is probably likewise an intrinsic property of being a car. But a fuel gauge or a rear-view mirror is not; at one time, cars were designed without either of these. They were still cars, and they would still be considered cars today.

I will not claim to be naming all the desiderata nor naming them the definitive way. After all, philosophers talk of concepts in various technical ways, and the various properties of concepts can likewise clearly be carved up in different ways[1]. Many of them overlap, and all of them inter-relate.

[1]For example, in a paper I wrote with Ron Chrisley, we offered the following list (not meant to be exhaustive): rational, articulable, recombinable (what I here call "compositional"), and under endogenous control (Chrisley and Parthemore, 2007b, p. 45). In Section 3.3.2, I will argue against articulability as an intrinsic property.

It may turn out that whether a property counts as intrinsic or extrinsic (or simply irrelevant) will turn on what perspective we happen to be taking at that time, not least whether we are considering concepts as we reflect upon them *as* concepts or as we possess and employ them non-reflectively. My goal, instead, is to provide a coherent and well-reasoned list that will apply equally well to concepts either as representations or as non-representational abilities, and that can best support the theory of concepts put forward in chapters Six and Seven.

In order to get the discussion started, I will offer a property baseline that most, though not all, philosophers of concepts can ascribe to, and build from there.

3.1 Evans' Generality Constraint

> In discussing the nature of our conceivings we have little enough to go on, but there is one fundamental constraint that must be observed in all our reflections: I shall call it "The Generality Constraint" (Evans, 1982, p. 100).

With his Generality Constraint, Evans is trying to set out what, to borrow a turn-of-phrase from Fodor (1998), might be called the "non-negotiable" properties of concepts and to be precise about what, at heart, makes structured thought structured. Fodor has his own, longer list of non-negotiables of course but, as with most of Fodor's positions, his list is somewhat tendentious, whereas Evans' Generality Constraint (for the most part) is not.

The Generality Constraint is deliberately minimalist. Remember that Evans' stated preference is to think of concepts not as representational objects but as non-representational abilities. Remember Frege's lesson: abilities are not things we explain so much as things we *use*. Evans continues:

> When we say that a subject's understanding of a sentence, *"Fa"*, is the result of two abilities (his understanding of *"a"*, and his understanding of *"F"*), we commit ourselves to certain predictions as to which other sentences the subject will be able to understand; furthermore, we commit ourselves to there being a common, though partial, explanation of his understanding of several different sentences (Evans, 1982, p. 101).

To wit, if the agent in question understands *Fa* and *Gb*, then it seems to follow that he must be able to entertain the propositions *Fb* and *Ga;* and, to the extent that *b* appropriately goes with *F* and *G* with *b*, make appropriate sense of them[2]. Furthermore, the agent should be taken to understand a whole host of other propositions of the form *F* is _____, *G* is _____, _____ is *a*, and _____ is *b*, in consequence of the understanding of *F, G, a,* and *b* being consistent across all their applications: i.e., it will be "possible for a subject to think of an object in a series of indefinitely many thoughts, in each of which he will be thinking of the object in the same way" (Evans, 1982, p. 104).

Some (e.g. (Beck, 2007)) have objected to the Generality Constraint on the grounds that it makes a metaphysical claim out of an empirically decidable one, so that, although it may be contingently true of all actual conceptual agents, nonetheless it is conceivable (and not *prima facie* self-contradictory) that some hypothetical conceptual agent might violate it and so, e.g., be able to entertain "John is happy" and "Susan is sad" but not "John is sad" or "Susan is happy". I

[2]Evans adds this caveat precisely to avoid the possibility of such arguably semantically uninterpretable constructions as Noam Chomsky's famous and oft-repeated "colorless green ideas sleep furiously".

will say nothing further here than that I shall take the burden to be on such a theorist to say how this is meant to work, and why the agent should still be considered a conceptual agent. The counterexamples offered by Beck apply only to what would normally be construed as non-conceptual content. I am not attempting to defend the Generality Constraint as a constraint on all cognition, only conceptual cognition.

Another critic, Charles Travis, objects to the Generality Constraint on the grounds that it requires structured thought to be independent of context, or, as he prefers to write, "sensitivity to surroundings" (Travis, 1994, p. 176). Although he does not address Evans' nod to "the categorical appropriateness of the predicates to the subjects" (Evans, 1982, p. 101), he presumably considers it a kludge. I think that Travis is fundamentally mis-reading Evans in thinking that the Generality Constraint is meant to be read as demanding that "to grasp any particular thing which was said, in particular surroundings, in calling some item **a**, F, one must be able to grasp what would be said, in any surroundings whatever, in calling any item whatever F, and what would be said, in any surroundings, in then calling **a** anything one could think about at all" (Travis, 1994, p. 188). In any case, Travis is explicitly *not* rejecting systematicity, and he is not concerned with productivity, which are the two properties I wish to derive as being at the heart of the Generality Constraint.

3.1.1 Systematicity

Concepts – or, *per* Evans' preference, conceptual abilities – exhibit systematicity, what Fodor (1998, p. 26) refers to as "symmetries of conceptual capacities". It would be strange to claim that a certain agent understands the proposition *X ys Z* even though it does not understand the proposition *Z ys X*. Indeed, as Fodor points out, conceptual understanding is systematic in a way that the world is not: it may well be the case that John loves Mary but that Mary does *not* love John. Fodor means this even much more broadly than the reference to symmetry would suggest, as (Fodor and Pylyshyn, 1988, p. 37) implies: "What we mean when we say that linguistic capacities are *systematic* is that the ability to produce/understand some sentences is *intrinsically* connected to the ability to produce/understand certain others" – and here there is no symmetrical qualifier. The systematicity of linguistic capacities, Fodor suggests, is parasitic on the systematicity of conceptual thought. In a similar vein, Prinz writes: "Saying that certain thoughts exhibit systematicity implies that our ability to have thoughts with certain intentional and cognitive contents carries with it the ability to entertain other thoughts with distinct cognitive and intentional contents" (Prinz, 2004, p. 14).

In the absence of any evidence to the contrary, it seems to many not just contingently but necessarily true that conceptual thought exhibits this methodical regularity.

3.1.2 Productivity

In similar fashion concepts are productive, so that a finite set of them can nevertheless give rise to an unbounded number of complex concepts and propositions.

Consider: the vocabulary of English, corresponding to the lexical concepts entailed by the language, is large – an unabridged English dictionary will have several hundred thousand entries – but certainly finite. Meanwhile a familiar and plausible claim, overheard at conferences, is that most of the statements of English being expressed at any particular time have never been constructed before. If this is indeed true, the reason is not just the simple combinatorial rules offered above, but

concepts' apparently unlimited capacity for compound (involving conjoined independent clauses), complex (involving subordinate clauses), and recursively defined structures (e.g., "the cat that the dog that the man rescued chased died") – in which ways English is in no ways unique. A repository of all the possible sentences of English would, for size, dwarf even Jorge Luis Borges' Library of Babel (2007), which after all imposed strict limits (on pages, lines, and characters) on the books in its collection. In analogous fashion, the number of propositionally structured thoughts[3] potentially thinkable must be, though not infinite (for reasons including time and mental resources, and constraints toward thinking certain thoughts and against thinking others), unbounded.

Note that, when Fodor writes about productivity, his focus is on generative capacities, while Evans' interest is as much or more in the ability of an agent to *understand* the propositions thus generated. Indeed, this is how, with respect to statements of a language, Chomsky himself uses the term: "...the ability of a speaker to determine, of a sequence of words he has never heard, whether or not it is a well-formed sentence, and what it means" (Chomsky, 2006, p. 81).

3.2 Further Core Properties

There are several other properties I will take to be intrinsic to the nature of concepts. The first is presupposed by the Generality Constraint, the second at least implied. The third is potentially controversial, and the last, though commonly held, *is* genuinely tendentious.

There are several further properties, often taken to be intrinsic to concepts, which I will argue are better understood as extrinsic. I will consider those in turn.

3.2.1 Intentionality

> Every mental phenomenon includes something as object within itself, although they do not do so in the same way. In presentation, something is presented, in judgment something is affirmed or denied, in love loved, in hate hated, in desire desired and so on (Brentano, 1995, p. 88).
>
> Since sentences... are, considered in one way, just objects in the world like any other objects, their capacity to represent is not intrinsic but is derived from the Intentionality of the mind. The Intentionality of mental states, on the other hand, is not derived from some more prior forms of Intentionality but is intrinsic to the states themselves. An agent uses a sentence to make a statement or ask a question, but he does not in that way *use* his beliefs and desires, he simply has them. A sentence is a syntactical object on which representational capacities are imposed: beliefs and desires and other Intentional states are not, as such, syntactic objects... and their representational capacities are not imposed but are intrinsic (Searle, 1999, p. vii).

The first property I will consider is the most obvious – so obvious, perhaps, that much of the time it is not explicitly stated. Concepts exhibit intentionality: i.e., aboutness. (Sometimes that aboutness is understood as consciously directed aboutness. Brentano made no such assumption, however, and neither shall I.) That is to say, they are about things, and the things they are about are, at least in the main, not concepts but, in Frege's language, objects. So for example a concept

[3] Recall from the last chapter that these need not be assumed to be linguistically structured.

of unicorn is about unicorns, regardless of whether (and in what sense) unicorns exist. There is no sense to be made of a concept that is not about anything.

In linguistic terms, this aboutness can be called *semantic content.*

It should be clear by this point that I wish to reject Searle's distinction between derived and non-derived content and the way he uses it to distinguish between two sorts of representations. *Pace* Searle, representations are necessarily conceptual, but concepts, though they are intentional, are not necessarily representational, precisely because they may lack the conscious direction I take as the hallmark of the representational.

For more or less the same reasons, I likewise wish to reject Frege's sense/reference distinction (1892), according to which there are two sorts of aboutness to consider: *what* a concept is about, and *how* it goes about being about it; in other words, one must account for conceptual content that cannot be provided solely by reference. So "the Morning Star" and "the Evening Star" refer to the same object (the planet Venus) but in different ways, by different Fregean senses or *modes of presentation*. One can possess the concept of "the Morning Star" and the concept of "the Evening Star" without realizing that they are co-extensive. On Frege's account, **UNICORN** has a sense (it refers in a certain sort of way) but not a reference (it does not, in fact, refer *to* anything).

As Prinz notes (2004, p. 6), Frege introduced this distinction in a discussion of language; but the distinction can be (and frequently is) taken to apply to concepts as distinguishable from language. Prinz by and large accepts Frege's account, though he prefers to talk of "cognitive content" rather than sense. Peacocke (1992) likewise makes the Fregean distinction.

Fodor, however, rejects it and insists that the conceptual content of beliefs is provided solely by reference (as required by his informational atomism). "Frege's problem... is problematic only if you work on the assumption that the *content* of a belief exhausts its contribution to the causal consequences of having it" (2008, p. 68). He allows that **UNICORN** is a genuine concept (see e.g. (2008, p. 151)) by taking a counterfactual approach: to possess the concept **UNICORN** entails that, were there any unicorns, the concept would lock to (all and only) unicorns. My own motivations are different, as I am claiming that only in the abstract can we make sense of reference as distinct from Fregean sense; we cannot step aside from experience to look beyond it.

On the account I will favour, **UNICORN** does have a referent, just one of a different kind from e.g. **DOG**, just as **TRANQUILITY** also has a different kind of referent from **DOG**. So far as we know, unicorns exist only in mythical worlds[4]– the stuff of imagination, of fairy tales and fiction; dogs exist there, too, but also exist in the physical world. The difficulty, if any, with **UNICORN** is that it appears to have the same sort of referent as **DOG** and it does not: one is a mammal; one is something else.

Or consider again the Morning Star/Evening Star. My forebears presumably considered these to be two distinct celestial bodies. I know (or rather believe, on the basis of expert advice) that they are the same celestial body, which is in fact a roughly Earth-sized planet between the Earth and the sun. By a process of conceptual change (see Section 3.2.4), what were two distinct concepts merged into one.

[4]...Spontaneous mutations (of which there are documented cases) and surgically altered circus goats aside.

3.2.1.1 The Role of the Conceptual Agent

I want to go one step further and say that, if concepts are necessarily about something, then they are, also, necessarily about something *to someone*: an agent, regardless of whether that agent is aware of possessing the concepts, perhaps of possessing any concepts at all (a point I shall return to shortly). An "agent" I will take to be any entity that interacts not just passively but actively with its environment: so all concepts are possessed by an agent, but not all agents possess concepts. Such a move is assumed by most if not all theories of concepts: concepts are discussed in the context of agents possessing and employing them. I am not aware of any theories that explicitly deny this.

To do so would be to break the essential link between concepts and cognition, unless cognition, too, is to be independent of agents; and, if anything bears the "mark of the cognitive" – to borrow a phrase from Frederick Adams and Kenneth Aizawa (2001; 2008) – concepts do. It follows that it makes no sense to talk about a concept of e.g. gold without some agent not just potentially but actually possessing it. If this is right, then there was no concept **GOLD** before there was an agent to possess it; and after the last agent who possesses the concept is gone, there will be a concept **GOLD** no more. In a slogan: concepts are never free floating. They require for their existence a context of both agent and environment.

One of the consequences of human cognitive evolution, to be discussed in Section 4.4.2.4, is the externalization of knowledge traces and, with it, the idea of conceptual knowledge as detached from any particular agents and perhaps *from any agents at all*. As useful, perhaps even necessary, as such an idea may be, however, it will remain a conceptual fiction.

3.2.2 Compositionality

> Concepts are the constituents of thoughts and, in indefinitely many cases, of one another (Fodor, 1998, p. 25).

As discussed in Section 2.4.1, concepts are compositional; indeed, productivity seems to require this, so that even the strictest anti-representationalist concepts-as-abilities advocates must, I believe, allow it. Any agent who possesses the concepts **BROWN** and **COW** *must* be able to entertain the concept **BROWN COW**; and any agent who possesses the concept **BROWN COW** *must* possess the concepts **BROWN** and **COW**.

It was noted in Chapter Two that there are different ways of understanding compositionality. Fodor's rationalism-inspired understanding differs in significant ways from the kind of compositionality understood by e.g. most connectionists, as well as other proponents of distributed representations[5]. Chalmers (1990) and Van Gelder (1990), among others, argue that Fodor is too quick to restrict compositionality to that which is syntactically explicit and seemingly transparent. Prinz (2004), likewise, makes use of a somewhat more liberal understanding of compositionality than Fodor's in arguing that proxytypes are compositional *enough*. I myself want to allow that, in many cases, concepts compose both upward and downward: i.e., they compose, and they can be decomposed. However, that concepts are compositional is not up for debate.

[5]Distributed representations are certainly not symbolic ones. They are probably best understood as a form of iconic representations.

3.2.3 Spontaneity

> If we want to be able to take it that the operations of conceptual capacities in experience impinge rationally on our thinking, as they must if they are able to be recognizable as operations of conceptual capacities at all, we must acknowledge that those rational relations fall within the scope of spontaneity (McDowell, 1996, p. 52).

Concepts are what Prinz (2004, p. 197) calls "spontaneous", borrowing (as does John McDowell) Immanuel Kant's (1999) distinction between receptivity and spontaneity, where both terms are being used in peculiarly philosophical ways. Receptivity is an organism's passive or pre-programmed encounter with its environment, while spontaneity is an organism's active (though, once again, not necessarily reflective) intellectual engagement with that same environment.

Stimulus-response behaviour is paradigmatically receptive: the choice for response comes from *outside* the organism: e.g. from its genetically inherited predispositions. Conceptually driven behaviour goes beyond what stimulus-response models can account for: roughly, where the same agent, presented with the same stimulus in the same circumstances, appears to make flexible responses – choices – at different times, based on an awareness of its own history. (Compare Albert Newen and Andreas Bartels: "To generalize our considerations so far, we do not need conceptual representations to explain any form of rigid behavior. Therefore, a minimal criterion for concept possession is flexible response behavior in one and the same stimulus situation" (2007, p. 286) or the "...possibility of the modification of a response in the light of additional information..." (2007, p. 287)).

Concepts are, on the other hand, paradigmatically spontaneous, being not simply given to the agent but rather actively developed or acquired by it: they are, in Prinz's words, under the agent's "endogenous control" (2004, p. 197).

No one, that I am aware of, would deny that concepts are typically spontaneous; but the possibility of innate concepts challenges the idea that concepts are necessarily so[6]. Fodor offers as "not seriously controversial" that "minds like ours start out with an innate inventory of concepts, of which there are more than none but not more than finitely many" (2008, p. 131).

What Fodor means by this must, however, be considered quite carefully. Although his radical nativism is often taken to entail that *most* concepts are innate – "Fodor's innateness thesis amounts to the claim that most of our concepts are innate, a result that virtually everyone finds patently absurd" (Kaye, 1993, p. 198) – Prinz notes that this is based on too quick a reading:

> ...One might wonder how evolution could have endowed us with some *particular* mental symbol that is predestined to track spatulas. If evolution set aside some symbol for this purpose, then it would have had to anticipate the invention of spatulas.... [However] if concepts are individuated by the properties that nomologically control them [as they are on Fodor's account], then to say that I have an innate spatula concept is not to say that I have some *particular* mental symbol in my head at birth.... It is only to say that I am disposed to enlist *some symbol or other* to serve as a spatula indicator. In other words, we are not born with spatula symbols; we are born with spatula detecting abilities... (Prinz, 2004, p. 230).

[6]...Unless those concepts are subject to spontaneous revision: see Section 3.2.4.

Fodor's more recent writings have attempted to straighten out many of the misunderstandings and make clear that what is innate is (or may only be) the conceptual abilities, not the concepts themselves. The acquisition process is, for Fodor, under the agent's endogenous control and hence, in Kant's terms, spontaneous.

> The central issue isn't *which concepts are learned*, since, if the (emended) LOT 1 argument is right, none of them are. Nor, however, is it *which concepts are innate*, if an innate concept is one the acquisition of which is independent of experience. Quite likely, there are none of those, either. Rather, the problem is to explain how a creature's innate endowment... contributes to the acquisition of its conceptual repertoire; that is, how innate endowments contribute to the processes that start with experience and end in concept possession (Fodor, 2008, p. 145).

The account put forward beginning in Section 7.2.3 will assume innate *proto-concepts* rather than innate concepts, precisely because the proto-concepts:

- Lack spontaneity.
- Are too few in number to be, of themselves, productive or compositional.
- Are not subject to change (see below).

3.2.4 Evolvability

> [Thesis of conceptual change:] ... Any concept that the subject possesses can change some of its properties while retaining its individual identity (Woodfield, 1994, p. 41).

Concepts change: on first blush, this may seem an unremarkable claim. My concept of gold, the argument goes, both is and is not the same as that of the ancient Greeks, since I know things about gold (or rather, believe on the advice of experts) that the ancient Greeks did not. Likewise my concept of gold both is and is not the same as my concept of gold as a child, when I had no clue to the existence of the periodic table. In both cases, enough has stayed the same to find commonality – justifying a common label; and enough has changed to deny strict identity.

On pain of vicious circularity, no structure can be entirely self-supporting. In order to get off the ground, conceptual change requires a considerable non-conceptual element: what Chrisley (2009) has called *non-conceptual conceptual change*, and which he compares to acquiring a skill like learning to ride a bike. Sometimes the right experiences – not structured in any conceptual way – are needed to push us into the new conceptual understanding. I will have more to say on the relationship between conceptual and non-conceptual content in Section 3.4.4.

Unlike systematicity, productivity, intentionality, compositionality, and spontaneity, I am not aware of anyone who specifically names evolvability as a property of concepts[7]. Nonetheless I take it to be assumed by anyone (such as e.g. Prinz (2004) or Gärdenfors (2004)) who allows beliefs to be partly constitutive of concepts. If beliefs can change, then so can concepts.

[7] McDowell writes of a "standing obligation to reflect" (1996, pp. 12, 40, 81), which implies something of what I mean. However, I have suggested already that not all agents we call conceptual agents will possess these reflective capacities; more in Section 3.3.1.

Conceptual change requires two things, as the quote from Andrew Woodfield implies:

- A core identity that stays the same.
- A peripheral identity that changes.

If the core identity itself changes, one no longer has conceptual change but conceptual obsolescence and replacement: not a changed concept, but a new one.

As with spontaneity, there is an important difference between the claim that concepts are *typically* subject to change and the claim that they are *necessarily* so. But some would deny that concepts – at least most concepts – can change at all, precisely by denying them any peripheral identity that *could* change. This includes those, like Fodor, for whom concepts are individuated solely by their extension, their content exhausted by their reference. My **GOLD** concept identifies the same stuff – gold – as that of the ancient Greeks; therefore, we have (precisely) the same concept; therefore, concepts do not change. (Of course, as noted in Section 2.3.2, what counts as "the same" is far from unproblematic. One might reasonably worry that what it means to "identify the same stuff" is precisely what a theory of concepts ought to address.)

Fodor allows that beliefs about gold can change, of course. But for Fodor, beliefs are not constitutive of concepts: concepts are constitutive of beliefs, but not *vice versa*. One could in principle possess the **GOLD** concept – because it reliably tracks gold – but have no beliefs about gold whatsoever.

Things get tricky because beliefs *are*, for Fodor, constitutive of stereotypes, and Fodor acknowledges that there is a significant relationship between concepts and stereotypes. (For sake of present discussion, stereotypes can be understood as prototypes; roughly, a stereotype is a preset idea of what a typical X is like, while a prototype *is* the typical X. Things are ideal when stereotypes and prototypes match.) In particular, he allows that learning a stereotype may well be a normal if not necessary stage in acquiring a concept. Allowing this, he says, "starts us on explaining why concepts are typically learned from their instances" (2008, p. 153).

One wonders if Fodor is not engaging in some philosophical sleight-of-hand, giving his readers good reason to believe that concepts are stereotypes/prototypes then rejecting that because, as he has consistently argued over the years, "concepts compose and stereotypes don't" (2008, p. 150). If Fodor is wrong about stereotypes/prototypes composing – and Prinz argues that he is – then by his own reasoning he is probably wrong about concepts not being stereotypes/prototypes, and if he's wrong about *that* then he has no argument against conceptual change.

Thomas Kuhn (1970; 1990) poses an entirely different challenge. Kuhnian paradigm shifts bring a *complete* changeover of concepts – what Kuhn terms a *Gestalt switch* – so that even when concepts go by the same name before and after, as they often do, it is a mistake to assume that they are related; in philosophical terms, they are *incommensurate*. My **GOLD** concept is not that of the ancient Greeks'; it is a new concept, because the discovery of the periodic table (among other things) caused a paradigm shift. I can understand the earlier concept – incommensurability need not imply incomprehensibility – but only by recognizing that it is a fundamentally *different* concept from the modern one. What goes for gold goes for much besides: "Pace the causal theorists of reference, 'water' did not always refer to H_20" (Kuhn, 1990, p. 4).

> In a similar vein, I've elsewhere pointed out... that the content of the Copernican statement, "planets travel around the sun", cannot be expressed in a statement that invokes the celestial taxonomy of the Ptolemaic statement, "planets travel around the earth". The difference between the two statements is not simply one of fact. The term "planet" appears as a kind term in both, and the two kinds overlap in membership without either's containing all the celestial bodies contained in the other (Kuhn, 1990, p. 5).

If all change were wholesale, radical change would, indeed, pose a problem for the sort of conceptual change I am describing – *if* there were no continuity, no sense in which the old and the new are still the same concept; but that would imply *complete* incommensurability, and Kuhn makes clear that that is not what he is talking about.

As Hanne Andersen and Nancy J. Nersessian note in analyzing the paradigm shift from Maxwellian to Einsteinian electrodynamics, "despite major conceptual change, there is still significant continuity between the Maxwellian and the Einsteinian concepts of field" (2000, p. S239) even though, on Einstein's account, the concept of "ether" has disappeared. As they interpret Kuhn, the radical change of paradigm shifts is complementary to *and continuous with* more incremental change.

Finally, there is the question of whether evolvability is intrinsic to or merely typical of concepts: i.e., whether *all* concepts are (at least potentially) subject to change. So long as beliefs are partly constitutive of concepts (i.e., part of being a concept is that one has certain beliefs with respect to it); and so long as it is part of being a belief – any belief – that it is potentially subject to change; then it will follow that concepts are subject to change as well.

In Chapter Seven I will suggest a yet stronger position: that concepts are in continuous if generally incremental motion through a constant process of acquisition, application, and change. This helps explain why the classical definitionists never found an untendentious definition: the definition is constantly shifting, both relative to an individual and relative to society. In any case, the image of concepts as unchanging entities is, like that of symbols, an idealization, an often useful approximation, and nothing more.

3.3 Extrinsic Properties

I want to provide a case that three additional properties should be understood as extrinsic: commonly associated but not essential, at least once an overly human-centric perspective is abandoned[8]. The three are closely related.

[8] For reasons to abandon such a perspective, see e.g. (Whitby, 2003).

3.3.1 Introspectibility

> To bring concept empiricism up to date, one must abandon the view that concepts are conscious pictures. Contemporary cognitive science helps in this endeavor by identifying a rich variety of highly structured, unconscious perceptual representations (Prinz, 2004, p. 139).
>
> But I do not detect any entity which could be called an occurrent abstract idea of Wolf. You may tell me that it is within my own mind.... I look into my own mind as carefully as I can, and still I do not find it (Price, 1969, p. 331).

Both classical definitionists and imagists, in their own ways, took concepts "just to be" the sort of things that we not only possess and employ but are actively *aware* of possessing and employing. I have suggested already in Section 3.2.1.1 that this need not necessarily be the case. Now it is time to explain why. At least two possibilities present themselves:

- A given conceptual agent might possess and employ a concept that, despite her being perfectly capable of reflecting upon it, never chances to come into her conscious awareness: the concept, though introspectible in principle, is not in practice. This is the situation Prinz appears to entertain when he goes on to talk of "highly structured, unconscious perceptual representations" that "can be used to form concepts" (2004, p. 139). Such "representations" might never enter the realm of conscious reflection; instead, they help structure the thinking that self-reflective agents *do* reflect upon.

- A given conceptual agent might possess and employ a concept that, despite her general capacities for reflecting on her concepts, she is, nevertheless, unable to reflect upon, even in principle. Recall the discussion from the last chapter of concepts as abilities: not all abilities *are* introspectible. "I don't know how I do it; I just do it!" One can ride a bicycle without knowing how it is that one can, which is to say, without knowing what knowledge *knowing how to ride a bicycle* depends on. Of course one might well be able to reflect on that ability; but if a given agent could not, we would not therefore say that she did not know how to ride a bicycle.

Both of those agents possess capacities for active reflection. However, it is a mistake to assume that, because we as humans are able to reflect upon our own concepts, that *all* agents we might want to consider conceptual agents must likewise possess that ability. It may be that, instead of allowing us to be conceptual agents in the first place, our reflective capacities rather transform what it means for us to be conceptual agents.

An agent that possessed mental content meeting all of the conditions in sections 3.1 and 3.2 but that lacked the reflective capacities of fully fledged self-awareness might still, appropriately, be termed a conceptual agent, if possibly to some diminished extent from humans. Two prominent writers in the "animal concepts" literature, Albert Newen and Andreas Bartels, write that "possession of concepts is not an either–or question, but something that develops gradually" (2007, p. 294), a position I wish to take as well (and will develop in Section 4.4 from an evolutionary perspective). In particular, they stress that "second-order conceptual capacities" implied by an ability for active reflection are not in any obvious way essential to first-order (non-reflective) conceptual capacities (2007, p. 289). In that light:

- A given conceptual agent might exhibit awareness of possessing a concept without exhibiting any awareness *of* that awareness: what, in Section 4.4.2.1, I call *implicit meta-cognition* and link to episodic memory and cognition, something that humans share with the other higher primates.
- Finally, an agent might conceivably exhibit signs of possessing and employing a concept without exhibiting any awareness – even implicit awareness – of doing so. Such an agent would lack, at least for the most part, even implicit meta-cognition. I suggest in Section 4.4.1 reasons for thinking that such an agent should still be considered as meeting a "conceptual baseline".

The opening quote from H.H. Price suggests another possibility: that concepts may not be introspectible *at all*, and that it is a mistake to assume otherwise. Still, if one accepts my position that beliefs are partly constitutive of concepts, then, because beliefs are in general introspectible, concepts should, in general, be at least partly introspectible: a typical but not necessary property of concepts and, hence, extrinsic.

Note that making introspectibility extrinsic is, though perhaps in tension, not in conflict with spontaneity being intrinsic, given how I have defined spontaneity as the agent's active, but not necessarily reflective, engagement with its environment.

3.3.2 Articulability

> The close connection of language to concepts in humans has seduced many into thinking that the two notions of language and concept cannot be disentangled (Allen, 1999, p. 39).
>
> Against the dependence of thought on language is the plain observation that we succeed in explaining and sometimes predicting, the behavior of languageless animals by attributing beliefs and desires and intentions to them (Davidson, 1987, p. 323).

By articulability I mean "potentially expressible in (spoken or written) language, by the agent possessing the concept". Articulability and introspectibility are closely related even though, for example, it is certainly possible (and consistent with the child development literature) that children are able appropriately to express certain concepts through language before they are able to reflect upon *them* rather than their objects. Likewise one may be able to reflect upon that which one cannot (at least as yet) articulate, perhaps because one lacks the appropriate language to do so.

Of course, no one wants merely to stipulate that language is necessary for concepts. One could argue that it simply is the case (subject to empirical evidence to the contrary) that no non-linguistic agents are conceptual agents – but then one must be careful not to be implying that, simply because we have no means directly to interrogate non-linguistic agents, they do not possess concepts. Unfortunately most of the time this reasoning is not spelled out.

I believe that McDowell takes concepts to require articulability in this sense[9]; Zoltan Torey goes further in making thought in general (2009, p. 46) and indeed mind itself (2009, p. 123) dependent

[9] I say this both because of the way McDowell describes language as bridging the divide between "mere animals" (who lack concepts) and humans (see e.g. (1996, p. 125) and all of Lecture 6 (1996, pp. 108-128)). The articulability requirement itself is given in passing in Lecture 1 (1996, p. 6). McDowell is willing to attribute proto-concepts to some non-human agents, meaning, as I do, mental particulars that do not meet all of the conditions on being concepts.

on language, and Davidson (1987) is similarly inclined. Although the later Wilfrid Sellars may well have moderated his views (de Vries, 1996), the early Sellars (1956) clearly took this position. Fodor does not explicitly endorse articulability in this sense (1998), but he does tie language and concepts quite closely together, so that most simple concepts end up being lexical concepts; and his language of thought hypothesis (LOT) effectively makes structured thought into a language of its own (1975; 2008).

I wish to present the case that concepts and language pull apart: that, even though we must unavoidably use language in order to talk about concepts, nevertheless we can, as when we recognize concepts as non-representational abilities, likewise see that concepts have a nature apart from language. This assumption is at the heart of Merlin Donald's account of cognitive-cultural evolution, which I will discuss in Section 4.4.

It is also an article of faith among the growing community advocating so-called "animal concepts": people like the aforementioned Newen and Bartels, or Colin Allen, who provided the opening quote. Of course, if those who would require language for concepts run the risk of being overly anthropocentric, then the proponents of animal concepts run the risk, as Newen and Bartels note, of anthropomorphizing animals (2007, p. 304). Nonetheless, they believe this bias can be avoided by careful application of certain empirically testable criteria; to wit:

- Evidence of an ability to derive general classes from specific instances.
- Demonstration of a flexible pattern of behaviour based on this ability, especially when confronted with novel situations.
- Demonstration of surprise upon making a mistake [10](Newen and Bartels, 2007, p. 291).

Of course, any behaviour, up to the most sophisticated human behaviour, could, in principle, be explained in terms of an inflexible stimulus-response mechanism. (Consider the standard philosophical example of a giant look-up table.) So to put these criteria another way, one should attribute minimal conceptual abilities to an agent when the most parsimonious explanation for that agent's behaviour is that, when presented with the *same* circumstances on *different* occasions, that agent makes different choices based on some awareness by that agent of its past experiences[11].

Newen and Bartels discuss some quite sophisticated experiments, originally presented in (Pepperberg, 1999), strongly suggesting that the parrot Alex, in all of the senses I have described as intrinsic to concepts, possesses concepts. They argue that the bonobo Kanzi possesses yet more sophisticated concepts in terms of ability to parse and to construct novel propositions, though they acknowledge that the empirical evidence is less solid, being based not on controlled experiments but on individual researchers' observations.

[10] A similar list may be found in (Allen, 1999, p. 37). Note that Davidson would explicitly deny this possibility for non-linguistic agents, in part because he believes that beliefs require (at least some) beliefs about beliefs – no first-order without second-order beliefs – and that such meta-cognition is only possible (or verifiable?) with language. At the same time, he allows that an agent without such meta-beliefs might be "startled" (1987, p. 326), so even he acknowledges the need for some continuum. I would accuse Davidson of over-intellectualizing matters.

[11] As Fodor puts it, in explaining why human beings definitely have mental representations and paramecia definitely don't: "Unlike paramecia, we are frequently implicated in primal scenes in which the behaviorally efficacious stimulus property... is nonnomic. Or, as I shall sometimes put the point in order to achieve terminological heterogeneity: the difference between paramecia and us is that we can 'respond selectively' to nonnomic stimulus properties and they can't" (Fodor, 1987, p. 10). Fodor's conditions in that paper for attributing mental representations to animals are, though worded quite differently, very similar in spirit to Newen and Bartels' list, or Allen's list, for attributing concepts.

I will have more to say about the relationship between concepts and language in the next chapter and in Chapter Seven. For now, allow that, if the animal concepts people are right, then concepts are not, necessarily, linguistically articulable for all agents who possess them. As with introspectibility, rather than allowing us to be conceptual agents in the first place, language may rather transform the ways we acquire, employ, and experience concepts: giving us new ways to learn concepts through discourse and reading, new ways to put them to use through speaking and writing. Language allows us to experience our most private thoughts as mediated by words when, as R.J.C. Burgener has written in support of Price (1957), there are a number of good reasons to think that many of our thoughts are *not* mediated by words. Consider the experience we all have had of having a word "on the tip of our tongues": we know there is a word for what we want to say, but we cannot find it.

3.3.3 Publicity

> Concepts are *public*; they're the sorts of things that lots of people can, and do, *share* (Fodor, 1998, p. 28).
>
> ...It will be useful to make a distinction that is fundamental to any discussion of concepts, but that is often neglected nonetheless. The distinction is marked in the questions "What is the (or our) concept of tree?" versus "What is Fred's concept of tree?" which deal respectively with social and individual notions of concept.... Philosophical arguments about animal cognition are also plagued by failure to heed the distinction (Allen, 1999, p. 35).

A too-quick reading of Fodor would understand him, by his publicity requirement, to likewise be adapting the articulability requirement as I have phrased it; but that would be a mistake. Fodor's target here is not an active, intentional, necessarily linguistically mediated sharing but a passive sharing, such that all agents who possess the concept X, however they come to possess it, *possess precisely the same concept*. Fodor's own articulability requirement I take to be, "potentially expressible in (spoken or written) language, *by some agent".*

I wish to reject even this milder form of the articulability requirement (which is the one I endorsed in (2007b, p. 45)), precisely because I want to call into question Fodor's assumption that concepts are (or should be considered as) *strictly* public, hence non-private, entities. I will reject Fodor's publicity requirement while leaving open the possibility of a much weaker version.

Some – e.g. (Woodfield, 1994) – talk of distinguishing private from public concepts as if to allow that they might be two, quite different, sorts of things: psychologists focus on the former, philosophers of mind on the latter. In a similar way, Edouard Machery (2009) distinguishes psychological from philosophical concepts, although he wants to go much further and eliminate concepts as any sort of unified category altogether.

I would prefer to talk about the private versus the public aspects of concepts, like two sides of a coin[12]. One is straightforwardly subjective, the other seemingly objective, or at least intersubjective. This is what one gets at when one distinguishes *the* concept of tree from Fred's concept

[12]I do not mean "private" in any "scary" sense, such as to be sealed off from public scrutiny. Such, I think, is the target of Wittgenstein's private language argument (2001, §256 ff.). It does seem conceivable to me that e.g. a linguistic agent might possess a concept (as a structured way of thinking about some aspect of her experience) that is strictly her own and not part of the shared experience of her society, and that she lacks the words even to begin to describe. But my arguments here in no way depend on this possibility.

of tree. For all of the importance of this distinction, many if not most discussions of concepts assume a position regarding it without making that commitment explicit.

Consider: my private aspect of **DOG** is partly constituted by my beliefs about dogs and is uniquely shaped by all of my dog experiences, including what I have read/heard/etc. about them. My public aspect of **DOG**, meanwhile, is partly constituted by my society's collective beliefs about dogs and shaped by its collective dog experiences. It is this latter aspect that is lexicalized, perhaps imperfectly, by the word "dog"; it is this latter aspect that coincides with Fodor's publicity requirement and his view on concepts more broadly. Although the two aspects more or less coincide, most of the time, nevertheless private relates to public in non-trivial ways, and from time to time the two come into conflict.

It is precisely with this private side of concepts in mind that Prinz calls for a relaxing of the publicity constraint: concepts need not be *perfectly* shared; they only need to be shared *well enough*. He writes: "If you and I agree about the most conspicuous walrus features, then we can understand each other when we use the word 'walrus', and we engage in similar walrus-directed behaviors" (2004, p. 158). When I speak of walruses, then your default assumption should be that I mean the word as you yourself would use it, that we are using it to refer to the same thing; assuming this yields good predictive advantages. Prinz continues:

> Communication is often imperfect. People generally manage to refer to the same objects and to associate many of the same features with those objects, but they do not necessarily associate all of the same features. My default proxytype for snakes may represent them as dangerous while yours represents them as harmless. My failure to understand why you do not flinch when I say there is a snake by your foot can be regarded as a very localized communication failure. Likewise, if we want to explain behavior by appeal to concepts, we often find situations where congruence is close but imperfect. I cower near snakes, and you do not. These minor discrepancies can be explained by saying that our concepts are similar but not exactly alike. Publicity does a better job of explaining communication and behavioral congruence if it admits of degrees (2004, p. 159).

The difficulties with similarity have already been noted (Section 2.3.2); perhaps Fodor is wise not to allow publicity to "admit of degrees". But Prinz may have a way out, by allowing (as he implicitly does) that public concepts have private aspects. The private aspects can (and probably must) be different even while the public aspects are (at least within a certain language community) the same. A similarity measure between individuals may not be needed: so long as the same concept (as designated by its public, lexicalized aspect) is employed *a clear majority of the time* in the same contexts by different individuals, it can be considered successfully shared.

Expectations of shared understanding can take us further than we realize, so that it is only when those expectations are rudely violated – e.g., when the conversation turns to flying walruses – that we start to question them[13]. In the worst case, communication may break down entirely: once shared understanding is questioned at *one* point, it may suddenly find itself open to questioning at *many* points. In the meantime, *assume all is well until proven otherwise* is an effective strategy.

[13]There are comedy sketches based around this very possibility, where each person in the conversation is involved in an entirely *different* conversation from the one the other is pursuing, without being aware of it.

Prima facie, a public/private distinction for concepts makes sense only with respect to enculturated linguistic agents. Though the non-linguistic agent may possess the propositional attitudes that form the core of our concepts in their private aspect, there is no public aspect to compare them against. Further, if the relation between private and public aspects is difficult enough with respect to the enculturated linguistic agent, it is unclear what comparison is to be made between the concepts of the non-linguistic agent and the public aspects of our concepts: e.g., in what way does the non-linguistic agent have a concept of dog or cat or tree, say, that is the same as the way we publicly relate to and talk about these things? In this (narrow) sense, Davidson is right: it is not at all clear what we are to make of "the dog's supposed belief that the cat went up that oak tree" (Davidson, 1987, p. 320).

3.4 From Concepts to Theories of Concepts

Before closing this chapter, it is necessary to move the discussion to a meta-level, to say a little more about not what a concept is but what a *theory* of concepts is – to set out *its* required properties and roles. Despite my own caveats on publicity, and the frequent complaint that philosophers and psychologists of concepts are talking past each other, still, I believe, most researchers have more or less the same target in mind.

Again, I do not intend this list to be definitive – the roles can be carved up differently, and theories of concepts can be put to different applications – but it should cover the main points in a fairly untendentious way.

3.4.1 A theory of concepts must explain conceptual agency.

A theory of concepts must say what it is to be a conceptual agent and distinguish, if appropriate, different sorts of conceptual agents. I will do so in Section 4.1.

A theory of concepts must likewise say what it is that allows us to distinguish an agent with concepts from one without: when, to borrow a turn of phrase from Daniel Dennett, it is appropriate to take the *conceptual stance*. This is *not* the same as saying what it means for an agent to be a conceptual agent. As Allen notes:

> It is important to be clear that the purpose... is not to provide a philosophical analysis of what it is for an organism to possess a concept. The question of when it is reasonable to attribute a concept to an animal is a distinct question from that of what it means for an animal to possess a concept..." (1999, p. 37).

Such a list of possible conditions has been set out in Section 3.3.2. This is more or less the one I will follow.

3.4.2 A theory of concepts must account for concept acquisition.

There are two sorts of concept acquisition one can talk about: phylogenetic and ontogenetic. Phylogenetically one can talk about the development of conceptual abilities in the species, which I will do in Section 4.4; or the possibility of innate concepts which, as I have already indicated, I will reject in favour of innate proto-concepts that, among other things, are not subject to revision and so not, as I have defined my terms, spontaneous. This leaves the burden of concept acquisition on ontogeny, where there are several possibilities:

- Concepts may be learned as e.g. *per* the developmental psychology accounts.
- Concepts may be "triggered", based on innate predispositions (see Section 3.2.3).
- Concepts may be surgically implanted, via some science fiction scenario, or acquired in some other unconventional manner, like a knock to the head.

I do not plan to pay any serious attention to the third possibility. I, also, wish to put aside the second possibility, on the grounds that it depends on the sort of nomic relations that I want to reserve only for the proto-concepts; all true concepts should have more complex relationships with their referents than nomic relations can support. So I will place my money on concepts being acquired by being learned.

Despite Fodor's radical nativist reputation (see Section 3.2.3), one of his "non-negotiables" on a theory of concepts is that "quite a lot of concepts must turn out to be learned" (1998, p. 27). As with his nativism, however, care must be taken to understand his notion of concept learning correctly, and not see him as immediately contradicting the earlier quote I gave from him. He emphatically does *not* mean "a process of inductive inference" (2008, p. 132), or, to put it another way, learning "' by abstraction' from experiences with their instances" (2008, p. 135): the position he is attacking; but that doesn't mean that concepts are not "typically learned from their instances" (2008, p. 153) – they are!

At the same time, the lesson he seems inclined to take from LOT (version one or two) is that "acquiring a concept from experience must be distinguished from learning it" (2008, p. 145); and what he means by "acquiring a concept from experience" is, more or less, the sort of "triggering" that Prinz ascribes to him, and which I wish to put aside. What he means by "learning" is harder to say, except that it does *not* involve explicit rule instruction or hypothesis testing. On this I can agree: there is no reason to think that, as a general rule, concepts are learned this way.

A lot has been written in the child development literature on concept acquisition, and one might do well to begin there. It is consistent with that literature to say that in all likelihood, children possess concepts (e.g. object permanence[14]) before they are able to express them; that they are able to express them through gesture before they are able to express them through language[15]; that they are able to express them appropriately through language before they are able to reflect upon *them* rather than their objects.

A promissory note in passing: concepts are not simply things that an agent collects; they have no meaning unless, at the same time they are being acquired, they are also being applied. A common mechanism for concept acquisition and application, taking some inspiration from the classical definitionist accounts (which also claimed a common mechanism), is provided in Chapter Seven.

[14] Jean Piaget, who coined the term, famously located this ability at age nine months (1954); more recent research (e.g. (Ballargeon, 1987)) has shown reliable evidence for an expectation of object permanence at less than half that age.

[15] "The first evidence of intentionality in children comes with pointing behavior.... Intentional pointing first emerges at about fourteen months, following a period during which children have learned to direct their gaze toward a point in space where their mother's gaze is fixed" (Donald, 1993, p. 171).

3.4.3 A theory of concepts must explain categorization.

> Categorical perception occurs when the continuous, variable, and confusable stimulation that reaches the sense organs is sorted out by the mind into discrete, distinct, categories whose members somehow come to resemble one another more than they resemble members of other categories (Harnad, 1990b, p. ix).
>
> Two general and basic principles are proposed for the formation of categories: ... The task of category systems is to provide maximum information with the least cognitive effort..... The perceived world comes as structured information rather than as arbitrary or unpredictable attributes (Rosch, 1999, p. 189).

To paraphrase Prinz (2004, p. 9), categorization is the epistemological counterpart to reference's ontology: rather than what we come to know/believe, it is *how* we come to know/believe. Categorization encompasses both the way that concepts carve up the world into categories (what Prinz calls "category production") – given some category, what attributes must or typically should members of that category possess – and the way they subsequently identify something as belonging or not belonging to a certain category. Categorization is, as Eleanor Rosch has famously described, not a chance process but one constrained by individual psychology and social dynamics.

Concepts themselves may be, and typically are, categorized along the lines of the things they are categorizing. So, for example, there is comparison of "action concepts" with "object concepts" (see Section 2.1). Likewise one can distinguish abstract (mental) concepts (**FREEDOM**) from concrete (physical) concepts (**STALACTITE**), and so on. I will have more to say about this in the next chapter, and quite a bit more to say in Chapter Six.

3.4.3.1 Concepts and Natural Kinds

> It is reasonable that our quality space should match our neighbor's, we being birds of a feather; and so the general trustworthiness of induction in the ostensive learning of words was a put-up job. To trust induction as a way of access to the truths of nature, on the other hand, is to suppose, more nearly, that our quality space matches that of the cosmos. The brute irrationality of our sense of similarity, its irrelevance to anything in logic and mathematics, offers little reason to expect that this sense is somehow in tune with the world – a world which, unlike language, we never made (Quine, 1969, p. 13).

Related to this is the question of how categories of concepts relate to categories in the world (presuming for the moment that there are any). It is standardly assumed by proponents of natural kinds concepts – I have in mind someone like Brian Ellis (2005) – that certain concepts (the *natural kinds* kind) "carve nature at its joints" (a phrase taken from Plato). That is to say, the categories recognized by concepts mirror categories that exist *independently of concepts* in the world. Natural kinds philosophers take their cue from W.V. Quine although, for all that he would not wish to endorse my own position, some things he says can be seen in some ways to support it.

If one accepts the argument made in Section 2.7 that one cannot stop being an observer – cannot stop being an employer of representations, cannot stop being a concept user, even for a moment, to get at what things in the world are in the absence of representations and concepts – then one will be, at best, skeptical of most natural kinds arguments. One will be inclined to believe, instead,

that categories are things that conceptual agents impose upon the world, and that, although the world severely constrains the ways we can usefully categorize (on which point I can happily agree with Quine), it does not provide the categories (and here he might agree with me). This is to replace the nomic relations of natural kinds concepts – the same kind of nomic relations I reject (for the most part) with Fodor's informational atomism – with a much more complex relation.

More to the point, however, and the real weakness of the natural kinds arguments I am criticizing, is that, as a general rule, they presuppose a realist metaphysics. If one accepts that starting point, then perhaps all is fine; but there is no compelling reason to do so, and I have already offered reasons why one might not.

At the same time, one need not be an anti-realist to question a too-simplistic mapping of conceptual categories to categories in the world. Indeed, Quine himself notes the problematic relationship between the notions of kind and of similarity, and suggests that, in a fully developed science, the notion of kind will disappear altogether (1969, p. 22).

3.4.4 A theory of concepts must relate concepts to non-concepts.

Concepts carve the world into those things that are concepts (paradigmatically cognitive) and things that are not. The divide they create is played out in many, seemingly unresolvable, oppositions: self/world, contemplation/experience, mind/body, spirit/flesh. On one side, more or less, are concepts; on the other, everything else. The enactive philosophy I introduce in Chapter Six stresses a continuity as underlying each of these conceptual distinctions. For now, it is enough to say a little about how a theory of concepts must address them.

3.4.4.1 Non-conceptual Referents of Concepts

Concepts, I have said in Section 3.2.1.1, belong to an agent. They are, in the main, about things – objects, happenings, properties, and the like – that are *not* of the agent and are not concepts, but of the agent's seemingly non-conceptual world. A theory of concepts can be expected, minimally, to explain the nature of the relationship between concept and referent, and the commonality (if any) between them.

What does it mean for **DOG** to be a concept of dogs (and not, say, cats); and can anything further be said about the relationship of **DOG** to dogs other than that they reliably co-vary? As an atomist, Fodor must say no: concepts are physically instantiated symbols, and that *is* all that can be said. On other accounts though, including prototype-based and other similarity-space-based accounts, there is typically an isomorphism between aspects of **DOG** and aspects of dogs – as I discussed in Section 2.4.2.

This does not mean, however, that concepts and referents need have much in common. "... Representations need not be similar to the objects they represent. What is important is that the representations preserve the similarity relations between the objects they represent..." (Gärdenfors, 2004, p. 109), borrowing a point from Shimon Edelman. What this means is: it is much less important whether **DOG** resembles a dog than whether e.g. the relationship between **DOG** and **CAT** is similar to the relationship between dogs and cats.

In any case, if the relationship between concept and referent is anything other or more than a bare nomic one, then a theory of concepts can be expected to explain not just the nature of concepts but something of the nature of their referents as well. I have already noted the tendency to classify types of concepts by types of referents. I will say more on this in Section 4.2.

3.4.4.2 Non-conceptual Content of Experience

> Content is non-conceptual just if it can be attributed to a subject without ipso facto attributing to that subject mastery of the concepts required to specify it (Bermudez, 2007b, p. 55).

The question arises whether our thoughts are structured, in part, out of stuff that does not meet (or does not fully meet) the conditions of being conceptual. Gareth Evans (1982) introduced the notion of non-conceptual content in discussing unconscious cognition, but contemporary discussions focus mainly around the content of conscious experience.

A theory of concepts can be expected to address whether there is such non-conceptual content, and, if so, what is its relation to conceptual mental content.

For non-conceptualists – people including Ron Chrisley (1996), Christopher Peacocke (1992), and José Luis Bermudez (2007a; 2007b) – conceptual sits alongside non-conceptual mental content, and both are typically, if not universally, present in experience. Non-conceptualists typically emphasize the continuity of the conceptual with the non-conceptual; whereas conceptualists like McDowell (1996) (if they acknowledge the non-conceptual as mental content at all – McDowell famously does not) see a strict distinction between the two. For non-conceptualists, conceptual and non-conceptual mental content are something like figure and ground: as figure and ground together form a picture, conceptual and non-conceptual content together constitute experience. For many conceptualists, on the other hand, experience only *is* experience to the extent that it is conceptual[16]. I will, in the discussion that follows, generally (though perhaps not always) side with the non-conceptualists.

3.4.4.3 Non-conceptual Foundations of Cognition and Life

Concepts as aspects of abstract, high-level, generally introspectible cognition reliably arise in the context of concrete, low-level, non-introspectible unconscious and sub-personal cognition. A theory of concepts has the burden to address how this happens as well as the relationship between levels of cognition more generally.

For people like McDowell and Fodor, there is a strict separation between levels. McDowell's "space of reasons" (1996) is complete unto itself. Likewise for Fodor as well as other functionalists, conceptually structured thought provides a functionality (strictly) independent of the low-level mechanisms that implement it: the same functionality can be implemented in many different ways.

To the functionalists, if an artefact exhibits conceptual functionality, say by the standards given in Section 3.3.2, it should not be denied designation as a conceptual entity simply because of how that functionality happens to be implemented; to do so would be "biological chauvinism" – an intuition

[16] I have Mike Beaton to thank for this phrasing.

that Searle famously challenges in (1980). While not outright denying the intuition, enactivists stress the continuity between levels of cognition: i.e., their lack of functional independence (see e.g. (Varela et al., 1991; Thompson, 2007)). Of course, many enactivists go further and make cognition and life co-extensive: see e.g. (Stewart, 1995; Ziemke, 2007). While otherwise following an enactivist line, I will, on this last point, remain agnostic.

3.4.5 A theory of concepts must relate concepts to each other.

On Fodor's informational atomism account, concepts exist entirely independently of each other (as noted in Section 2.4.1). On many if not most other accounts – clearly those like Davidson's (1987) that tie concepts closely to propositional attitudes (as I also wish to) – it is not possible to have any one concept without having many other concepts.

Most concepts are about things that are not, themselves, concepts. But some, like the concept of a concept itself, explicitly reference (other) concepts. These second- and higher-order concepts are sometimes considered as discrete levels, sometimes as rough positions along a continuum. I will have more to say about them in Section 4.1.1.

3.4.6 A theory of concepts must be empirically testable.

Finally, in keeping with the spirit of contemporary analytic philosophy, a theory of concepts should, wherever practical, make empirically testable claims. Of course, many things in this field are not, for practical reasons or by their nature, directly testable – how does one prove that a concept is or is not a proxytype? – but some of them can still be addressed indirectly. So for example, one could build a computational model of a theory that can be run through various simulations; an iterative process of theory-model-implementation-theory may be the best means for pushing the theory forward.

In a similar vein, one could build an application for helping users externalize a portion of their conceptual domain and examine it for e.g. consistency. The ease of use and subjective sense of naturalness (or lack thereof) whereby a test subject engages the application would provide indirect but very useful evidence for or against the theory, as would the ability of the externalized domain to make predictions about the subject's "private" conceptual domain. A proof-in-concept of such an application is presented in Chapter Eight.

3.5 Conclusions

I have attempted, in this chapter, to set out a case for certain properties being intrinsic to concepts (roughly, individually necessary and jointly sufficient), and for others to be merely typical but, in certain circumstances, not necessary. In doing so I have sought to provide the space in which, in Chapter Four, to emphasize the continuity of human with non-human cognition, rather than presuming (as others seem to do) the uniqueness of the human. The idea that only humans possess or *could* possess concepts becomes restricted to the idea that only humans possess language[17]. Indeed, even *that* position may be subject to further qualification, reducing it to the truism that only humans possess human language, given the rich ways that many non-human animals communicate.

[17] Davidson usefully agrees with me that the matter is not ultimately empirical. "... The question is what sort of empirical evidence is relevant to deciding when a creature has propositional attitudes" (1987, p. 317) – which, like me, he wishes to make constitutive of having conceptual thought.

At the same time, this is in no way to trivialize the important ways in which human language *is* different from the communication systems in other animals, not least being (the lesson I take from Chomsky) our capacity to entertain recursively defined propositional structures – even though our capacity to construct and understand such structures is exceedingly limited (see Section 5.2.1.3). The lesson I take from Donald is that, in order properly to appreciate what *is* uniquely human, it is necessary, first, to come to terms with what seems to be uniquely human and is not.

In listing properties, I have not attempted to be exhaustive; if the arguments in Section 2.6.1.2 are anything close to being right, then the intrinsic properties of concepts should *not* form a discrete set; while the set of extrinsic properties is potentially open ended. Rather, the attempt has been to sketch a coherent picture roughly in keeping with folk intuitions and philosophical inquiry. The most important conclusion of this chapter is that, at least in the human case, at least most concepts have a public and a private aspect, and the mapping between the two is less than straightforward. Although this distinction is often neglected in the literature, it is critical to understanding historical and contemporary theories of concepts, and it helps for example to place Frege's sense/reference distinction in a different light: Fregean sense has more to do with the private aspect, reference with the public[18]. The outcome of the extrication is the first step toward an enactive theory of concepts, the unified conceptual space theory of chapters Six and Seven.

The chapter concludes by drawing together ideas from the current and previous chapter to say in more detail what constitutes a theory of concepts and what it should be expected to account for. As itself a conceptual activity, a theory of concepts takes a meta-perspective on concepts, turning attention away from concepts' focus on the world to the concepts themselves.

[18]...With, needless to say, no need for a strict dividing line between the two.

Chapter 4

Concepts in a Context of Agents, Referents, Use

In the novel ***Johnny Got His Gun***, the protagonist, a victim of the battlefield, is deaf, dumb, blind, and cut off from nearly all external stimuli. Much of the novel takes place within his mind, as he is caught between fantasy and reality (Trumbo, 2007). Even though he is a prisoner in his own mind, nevertheless his body continues to interact with its surroundings in rich ways – some of which eventually make it through to consciousness.

For most of us, though, the world is constantly at hand, whether we are paying attention to it or not. We are embedded in a particular environment and embodied in a particular form; a sudden change to either of these *will* get our attention.

Just as we exist within an environment that helps to define us – pushing against us even as we push against it – so, too, concepts exist within a context of agents and applications that help to define them. Section 3.4 set the stage for this chapter by talking about the way that concepts and non-concepts fit together like figure and ground, and by suggesting that concepts can be classified according to:

- The different sorts of agents who possess them, by virtue of different cognitive abilities (Section 4.1).
- The different sorts of things they refer to (Section 4.2).
- The different uses to which they are put: i.e., the different roles they play (Section 4.3).

Of course, we do concepts an injustice if we fail to consider them in a context that is at once spatial *and* temporal. In particular, we need to consider, at least briefly, the development of concepts, both in the individual agent (ontogenetically) and over the development of the species (phylogenetically). Doing so will give us a better appreciation of the ways that, as I argued in Section 3.3.2, concepts and language pull apart; they will also give us a non-mysterian view of how language could arise for the conceptual agent in the first place (rather than requiring, as on e.g. Davidson's (1987) account, for language to be in place in order to have any thoughts).

Finally, just as concepts arise, are possessed, and are employed in context, so, too, as meta-conceptual entities, must theories of concepts. Though the aim generally is, and probably should

be, to provide the broadest possible theory, nonetheless the particular questions that are being asked and the particular applications intended will, inevitably, shape the theory.

4.1 Types of Concepts by Types of Agents

> The phrase ["having a concept"] itself is ambiguous. We say of a person who is able to talk intelligently about past, present and future events, read the clock, *etc.*, that he has a concept of time. He knows what time is; he recognises it, so to speak, when he meets it; he can ask "What is the time?", and can understand this question and answer it when asked. But there is also a sense in which he may not have a concept of time; like St. Augustine he may not be able to answer the question "What is time?" It is this second sense of "having a concept" that the upholders of what we might call the reflectionist position have in mind. For them to have a concept of X is to know what it is for something to be X and not just what is an X. It may seem rather odd that we should know what is an X without knowing what it is for something to be X, but for the moment we have the example of the concept of time to remind us that, odd though this may be, it is a fact of experience. For the sake of convenience I shall call concepts in the first sense – concepts which enable us to talk about and ask questions about *the* time – first-order concepts; those used in the second sense – concepts about *time*, etc. – I shall call second-order concepts (Barrett, 1962).

When we talk about the concept of time, there are, as Cyril Barrett so nicely shows, (at least) two distinct meanings we could have in mind. An agent might perfectly well (we might even say fully) understand the question "what time is it?" yet not be able to address the question "but what is time?" That is to say, she has *a* concept of time, but she has no means to address *the* concept of time: her own, or anyone else's. The second question, and not the first, requires the ability (to some greater or lesser extent) to reflect on time *as* time, to turn one's attention away from the application of time in practical matters to (the nature of) time itself. The agent's failure to address the second question might be a lack of a certain level of conceptual abilities or reflective capacities; or it might only be that the agent has had no prior reason to reflect on her concept of time, being sufficiently occupied just with using it.

Section 2.7 introduced a critical distinction between possessing and employing concepts non-reflectively and reflecting on concepts as concepts, while Section 3.3.1 raised the possibility of an agent possessing the first ability but lacking the second: i.e., lacking awareness of itself as a conceptual agent, regardless of any other reflective capacities it might possess. Let us call the two types of conceptual agents the conceptually reflective agent and the conceptually non-reflective agent.

The conceptually non-reflective agent could still possess (in Barrett's first sense) many if not most concepts one might name. To the external observer, such an agent would possess conceptual abilities with no need to attribute to that agent any conceptual representations.

However, there should be certain concepts that will be, in practice and in principle, beyond the conceptually non-reflective agent's reach. I suggest three inter-related categories.

4.1.1 Second- (and Higher-) Order Concepts

> Second-order concepts are concepts of concepts. They take first-order concepts for their objects, and clarify them (Barrett, 1962, p. 129).
>
> Let us start by adapting and endorsing Barrett's classification of "first order" and "second order" concepts. By this he divides concepts into those whose existence is in some way self-evident, and which are recognized at once when encountered, and, secondly, those which take first order concepts and clarify them. These second order concepts are those which are arrived at by reflection and abstraction. They are, in a sense, "concepts of concepts". Put rather more simply, having a first order concept means "knowing an 'x' when you see one", whilst having a second order concept means "knowing what is entailed by being an 'x"' (Eyles, 1973, p. 150).

Let me define a first-order concept as a concept of a non-concept (or, as will be useful later, a zeroth-order concept), a second-order concept as a concept of a first-order concept, a third-order concept as a concept of a second-order concept, and so on. More simply, one can talk of first-order concepts (concepts of non-concepts) and higher-order concepts (concepts of concepts, concepts of concepts of concepts, and so on). This requires a short digression by way of explanation.

The first description in particular suggests a rigid hierarchy in the spirit of Bertrand Russell's theory of types (1908). But a theory-of-types approach explicitly bans self-reference (of any kind) since no item on level n can refer to another item on the same (or any higher) level. As Douglas Hofstadter neatly describes, a theory-of-types approach would render language virtually unusable:

> A rather matter-of-fact sentence such as, "In this book, I criticize the theory of types" would be doubly forbidden in the theory we are discussing. First, it mentions "this book", which should only be mentionable in a "metabook" – and secondly, it mentions *me* – a person I should not be allowed to speak of at all (Hofstadter, 2000, p. 22)!

If language without self-reference is unusable, then one should, at the least, be wary of attempting a theory of concepts without self-reference. Among other things, this would disallow the concept of a higher-order concept or the concept of a concept itself. It would also establish an unbridgeable divide between concepts in general and one's own concepts.

But allowing self-reference raises problems of its own, which is precisely the reason, with respect to set theory, Russell was so anxious to avoid it[1]. Paradoxes – or, as Russell preferred to call them, contradictions – quickly arise. To understand why, consider the following paraphrase of Grelling's Paradox (itself a variant of Russell's Paradox) into conceptual theory[2].

An appropriate description of higher-order concepts is *potentially self-referential*[3]. An appropriate description of first-order concepts is *not potentially self-referential*. Just as **HIGHER-ORDER**

[1] This is the same reasoning, I believe, behind Frege's absolute distinction between concepts and objects, and the source of Frege's dispute with Benno Kerry.

[2] Grelling's Paradox sorts adjectives into autological (self-descriptive): e.g., *polysyllabic, short, unhyphenated*; and heterological (non-self-descriptive): e.g., *monosyllabic, long, hyphenated*. The question then is, is *heterological* itself heterological or autological? In similar fashion, Russell's Paradox sorts sets into those that contain themselves as members and those that do not, wherein the question becomes: is the set of all sets that do not contain themselves as members a member of itself?

[3] The "potentially" qualifier is needed, of course, because not all higher-order concepts need refer, either directly or indirectly, back to themselves. All that matters, as we will see, is that some do.

CONCEPT and **FIRST-ORDER CONCEPT** are, themselves, concepts, so, too, are their corresponding labels: **POTENTIALLY SELF-REFERENTIAL** and **NOT POTENTIALLY SELF-REFERENTIAL**. In both cases, concepts must, if one accepts the law of the excluded middle, be either one or the other.

The question becomes: is the concept **NOT POTENTIALLY SELF-REFERENTIAL** *itself* a potentially self-referential concept (hence higher-order) or a *not potentially self-referential* concept (hence first-order)? If it is a potentially self-referential concept, then, precisely in the case where it is self-referential, it becomes a *not potentially self-referential* concept. But if it is a *not potentially self-referential* concept, that means that it is potentially (indeed, actually) self-referential.

One could attempt to separate higher-order concepts into *potentially self-referential higher-order concepts* and *not potentially self-referential higher-order concepts*. But one has not solved anything, only pushed the problem up a level.

This is why, I think, Eyles qualifies "concepts of concepts" with "in a sense". Although it will often be useful, in the discussions that follow, to treat concepts as falling onto discrete levels of zeroth-, first-, second-, etc. order, it should always be remembered that those levels are an imposition on what should probably be understood as an underlying continuum. The problem cases arise precisely where one draws the uncrossable lines.

Bottom line: there will always be some concepts that will be, at least most of the time, unambiguously first-order (**DOG**), others umambiguously and explicitly higher-order (**CONCEPT**). But others (**NOT POTENTIALLY SELF-REFERENTIAL**) will not be classifiable on one side of the line or the other – *wherever* one draws the lines (by simple extension of the argument above), so long as one allows self-reference (and at least implicit universal quantification). Still others, like **DOG**, while seemingly first-order, have an implicit higher-order aspect. After all, borrowing a page from Barrett, we can distinguish a question like "have you seen the dog?" from "what does it mean for something to be a dog?" or "what is the essence of dog-ness?". This is even more straightforwardly true with abstract concepts like **TIME**, precisely because abstract concepts are at least implicitly higher-order; see Section 4.1.2.

The difficulty is that requiring concepts to fall on one side of a line or the other is, as noted in Section 3.4.3, itself a conceptual distinction. As conceptual agents, we cannot step outside of our conceptual schemes to resolve the matter. Instead, we see things *now* the one way, *now* the other. Where conceptual understanding imposes a binary distinction, logic (in particular, the presentation of paradox arising from self-reference) suggests an underlying continuum. The thesis put forward in the next chapter is that paradox arises precisely where we attempt to see beyond this and similar binary distinctions imposed by our conceptual understanding, and cannot.

More will be said about self-reference in Section 5.2.1; for now, end of digression. Clearly some agents – notably humans and arguably higher primates[4] – possess higher-order concepts. Conceivably other agents – arguably the corvids – possess concepts but only the first-order sort. At least, the possibility of an agent possessing only first-order concepts is *prima facie* coherent. Likewise, there is the possibility of identifying a certain agent's concept of x as being only first- and not

[4]Higher primates have been shown to pass the mirror self-recognition test; see Section 4.1.3. Of course, the interpretation of these experiments has been hotly debated: e.g., by Torey (2009, p. 90).

second-order – so long as that agent is not oneself! Of course one could, and probably should, conclude that one possesses some concepts that are only first-order in this way. But to observe them as such is immediately to make them higher-order.

4.1.2 Highly Abstract Concepts

> Now the first hypotheses [sic] to be considered is that there are, in addition, certain *third order concepts*. For whereas first order concepts are concerned with things that exist, and second order concepts may be concerned with things that exist, or with qualities and relationships, third order concepts are concerned entirely with qualities and relationships. Moreover, they are concerned with the kind of qualities and relationships which have no existence in themselves, but subsist entirely in the realm of being (Eyles, 1973, p. 150).

The argument against a strict separation between levels and in favour of a continuum is reinforced by Eyles' discussion of third-order concepts. Though the implication of third-order concepts as *concepts of concepts of concepts* is lurking, Eyles does not reference it directly. Rather his interest, as Barrett's, is with increasing abstraction away from the practical and immediate to the theoretical and distal; from concepts of the concrete to concepts of abstractions, to concepts of *abstractions of* abstractions. (It is not correct, however, to say that higher-order concepts are abstract and first-order ones are not: "It should be noted that second-order concepts are concepts of abstractions (and not as they are commonly called, merely abstract concepts, since all concepts are in fact abstract)" (Barrett, 1962, p. 130)). So Eyle's "third-order" concepts should better be expressed as *highly abstract concepts.*

To return to the earlier discussion, what the conceptually reflective agent possesses and the conceptually non-reflective agent lacks is a capacity (along a continuum) to step back from the immediate / the present-at-hand / the world. Probably some nascent hint of this is present already in the conceptually non-reflective agent (see Section 4.3), but the conceptually reflective agent takes things much, much further. I think this is how Donald should best be understood when, writing about apes (perhaps the cognitively closest animals to humans), he say that "their lives are lived entirely in the present. . . "(1993, p. 149).

The paradigmatically concrete can be captured in a picture (or, perhaps, a sound, smell, or tactile sensation). The paradigmatically abstract resists any such easy illustration and can typically only be approached in a roundabout way through words. A bird can be pictured, a scream can be heard, yes; but what does contentment look or sound like? The best one can offer, perhaps, is a metaphor: "a kitchen full of baking bread" or "the sound of running water".

The conceptually non-reflective agent's conceptual world may be expected to centre on, or simply be limited to, such present needs and dangers as food, shelter, predators, comforts and discomforts, none of which require *prima facie* any high degree of abstract thought or capacity for reflective awareness. What need has such an agent for such abstract concepts as democracy, *deja vu,* or altruism – or the concept of a concept itself, as paradigmatically abstract and higher-order as one gets?

4.1.3 The Concept of Self-as-Myself

> It does not follow that a person who knows how to use the first-person pronoun has a concept of self. He certainly has a concept of I or of myself, and from this, by a process of reflection, may arrive at a concept of self; but (in itself) the ability to use the first-person pronoun is not evidence for somebody having a concept of self (Barrett, 1962, p. 129).

It is time to consider what *self*-reflection and *self*-reference imply for the conceptually reflective agent: namely, a concept of (agent as) self, or rather, a particular concept of (agent as) self. Whether brought into the foreground or pushed into the background, the "I" is always present for the agent capable of asking the question, "what is a concept?".

The conceptually non-reflective agent, by virtue of being focused on the world, is not focused on her*self*. Her attention is other-directed and "outward", not self-directed and "inward"[5]. Of course, self awareness may, again, be along a continuum, and that is what I take Barrett's seemingly paradoxical point to be implying.

Unlike higher-order concepts and highly abstract concepts, the concept of self is in a category all of its own. It is unique among all a conceptual agent's repertoire of concepts in that it points back to the agent. As the agent is the master of her concepts, so, in this way, the **SELF** concept may be seen as the master of all the other concepts.

Care must be taken here. Even many agents to whom we would not attribute conceptual abilities at all make an effective self/other distinction. What would be the survival chances of a predator that made no such distinction in determining what to attack and eat? Indeed, as Dennett has pointed out, *all* organisms that show a habit of self-preservation must make some such distinction (see the discussion in (Dennett, 1991a, p. 174)).

Thomas Metzinger opens his book ***Being No One*** with the provocative assertion that "...No such things as selves exist in the world: Nobody ever *was* or *had* a self" (2003, p. 1). There is a curious problem deciding to what the concept of self or the "I" refers. It could be:

- The biological organism (I_1); or
- The self-aware sense of mental presence (I_2); or
- The name (if any) or other simple self-description an agent gives itself (I_3);
- Or all three, conflating the three together (as we often do), even though they are, conceptually, distinct.

One can deny, as Metzinger does, the ontological existence of self, but not the role that the concept of self plays. In developmental psychology, the self/other (self/non-self, self/world) distinction is seen as foundational to all the other concepts we as humans acquire (see e.g. the discussion in (Zachar, 2000, p. 144 ff.)). What Metzinger rightly points out is that the self cannot be what it is often taken to be: a matter of strict identity between concept and referent. For if the self-reference were taken at face value, then the part (the "I", an aspect of the agent) would swallow the whole (the entire agent): a logical impossibility. Putting this another way: strict self-reference would

[5] "Outward" and "inward" are in scare quotes for reasons discussed in Section 1.5.

have the "I" referring to itself and referring to itself *referring to* itself and so on, *ad infinitum*. This is the counterpart to the Grelling-like paradox discussed in Section 4.1.1: in that case, one arrived at an eternal oscillation between two untenable positions; in this case, one faces an eternally receding target. Again, more will be said in Section 5.2.1.

Of course, as Barrett points out, not every conceptual agent with a concept of self – and this might be all agents one considers conceptual agents – necessarily has such a sophisticated concept of self. The animal-concepts people suggest that many animals might qualify as conceptual agents; but only a few animals, besides humans, can pass the mirror test.

A simple conceptual agent may have only what I have labeled I_1 above: a first-order concept of the organism "as a whole", without distinction of body or mind or anything else. This is close to Antonio Damasio's notion of the *core self* (2000).

It is quite distinct from the higher-order self-as-myself that most humans entertain, and which requires, or perhaps creates, the body/mind distinction. Who does the "I" who thinks "I" think that "I" is? This is Damasio's *autobiographical self* (2000).

This is the self as strongly distinct from all the other agents in the world. This is the self that is, if we are careful not to confuse the metaphor with the reality, the homunculus sitting in his Cartesian theatre of the mind, controlling the shell of an organism in which he sits and observing all that it observes[6]. Rather than its target being *this* thought or *that* thought, as with most other higher-order concepts, its target is the agent entertaining those thoughts, or, more accurately, the agent's *conceptual understanding* of the agent entertaining those thoughts. That is to say, the referent of this concept of self is *not* the self itself, but the agent's concept or model of that agent, an idea which Metzinger would surely find congenial.

If the self-as-myself (I_2) is already very abstract, then the third notion of self I have introduced (I_3) is very rarefied indeed. Like any good symbol or label, it is little more than a placeholder.

[6]Of course, if the metaphor were the reality, the homunculus would invite another homunculus to be observing him, and so on in an endless regress (see for example (Harnad, 2006)), just as I have described above. Homuncular regresses are generally, indeed, things to avoid; but one should not ignore or abandon the power of an idea just because of how it might be misapplied.

4.2 Types of Concepts by Types of Referents

4.2.1 Physical vs. Mental (or: The Mind/Body Problem)

> The mind-body problem, so construed persists in philosophy because of two intellectual limitations on our part. First, we really do not understand how brain processes cause consciousness. Second, we continue to accept a traditional vocabulary that contrasts the mental and the physical, the mind and the body, the soul and the flesh, in a way that I think is confused and obsolete (Searle, 2002, p. 57).
>
> Twenty years ago, emotions, qualia, and "raw feels" were held to be the principal stumbling blocks for the materialist program. With these barriers dissolving, the locus of opposition has shifted. Now it is the realm of the intentional, the realm of the propositional attitude, that is most commonly held up as being both irreducible to and ineliminable in favor of anything from within a materialist framework (Churchland, 1981, p. 67).

First of all, concepts can be sorted according to whether they refer to physical things or mental things. There are concepts of the mental and concepts of the physical. But some – notably the Churchlands (Churchland, 1981, 1989) but also Searle – want to do away with any such familiar distinction as antiquated, either an unfortunate accident of our cultural upbringing or a consequence of a folk psychology as deeply flawed as the theories of alchemy or an Earth-centric universe. To the contrary, I will suggest that the distinction is useful and needs to be maintained, even while denying, along with the Churchlands and Searle, that there is any substantive mind/body "problem".

Much philosophical blood has been spilt over the question of dualisms of one kind or another, and one might be excused for thinking all dualisms bad, or at least things to be avoided. At the same time, monism, as an absolute metaphysical position – the view of reality as a single, unified whole, a conceptual singularity – is a difficult thing to hold to. Dualisms, at least of the right kind, are useful, if not in fact conceptually necessary, an idea I freely borrow from David Papineau (see e.g. (2006)).

Prima facie, mental and physical are substantively different things, with different properties inhering in each. One has an intrinsic subjective element, the other seemingly none. One is spatially (and temporally) extended, possessing properties like length, width, and height; the other is not.

Certainly *Cartesian substance dualism* – the idea that mental and physical are ontologically separate substances – is out of fashion, if it was ever truly in. Nobody, it seems, wants to support substance (Cartesian) dualism – David Chalmers, accused, denies the charge (Chalmers, 1996) – and it's far from uncontroversial that Descartes supported the position that came to bear his name (Baker and Morris, 2002; Oakeshott, 1991, p. 22).

The many varieties of *property dualism*, on the other hand, say that mental and physical are not two kinds of substances but two kinds of properties. Propositional attitudes and consciousness, for example, are mental properties of physical objects – brains – just as volume and mass are physical properties of those same objects. Such distinctions are reminiscent of John Locke's account of primary and secondary qualities (Locke, 2004), where primary qualities are meant to inhere in physical objects in an observer-independent way, while secondary qualities do not: they are, as it

were, "in the eye of the beholder". Searle (2002), who has been accused of being a property dualist, rejects property dualism as amounting to nothing substantially different from substance dualism.

Per physical monism/physicalism, I believe it a mistake to see physical and mental as being in opposition to each other: i.e., two instances of one genus, be it of substance or property. I wish to reject both substance dualism and property dualism, and adopt, instead, a position with elements of Papineau's *conceptual dualism* (2006). Let me call my position *perspective dualism*. Mental and physical are neither different substances nor different properties but rather different perspectives along the same continuum I discussed in Section 4.1: zeroth-order concepts are straightforwardly physical (the things they are about *look* physical), sufficiently higher-order concepts are straightforwardly mental (the things they are about *look* mental). Everything in between is somewhere along the continuum. At the same time, these perspectives are not things we can set aside or move beyond; they are imposed on us by the nature of conceptual thought. There is no useful distinction to be made between primary and secondary qualities because, as I will argue in Section 4.2.2.3, all properties (and indeed all concepts!) are abstract and mental; all are observer dependent.

Pace many physicalists – and certainly the eliminativists like the Churchlands – I think it is conceptually irresistible not to blur the distinction between mental and physical, because properties like mind often behave as though they were objects of substance, precisely because of this continuum. Physicalists typically take the physical world as given and unproblematic, the mental as what is problematic – a position aggressively challenged by Karl Popper (1982). I consider it a mistake to take either mental or physical as unproblematic: caught within our limited perspective, there is an unavoidable tension between the two. If there is a mind/body problem, it is merely a conceptual one, for underlying the binary opposition is a continuum between mind and body as there is between mental and physical.

4.2.2 Objects vs. Action/Events vs. Properties

Object concepts, action/event concepts, and property concepts refer to noun-like things or "objects", verb-like happenings, and adjective-like properties, respectively[7]. With respect to the mental-physical continuum, object concepts and action/event concepts will be more toward the first-order and physical, property concepts more toward the higher-order and mental. These distinctions will prove central to the unified conceptual space theory put forward in Section 6.2.

[7]This relation, between conceptual and linguistic categories, has been widely noted (see e.g. (Hemeren, 2008) for the first two and (Gärdenfors, 2004, p. 60) for the last).

4.2.2.1 Objects

> Much of the research dealing with the connection between what we see and the subsequent lexicalization of our percepts into concept hierarchies has been mainly addressed from the perspective of the object-noun relationship (Hemeren, 2008, p. 55).
>
> For well over one hundred years, thinking about the representation of object concepts in the brain has been dominated by sensory-motor property models.... The central idea is that object knowledge is organized by sensory features (e.g., form, motion, color) and motor properties associated with the object's use... (Martin, 2007, p. 25).

Objects, toward one end of the continuum, will simply be physical objects; toward the other end, they will bleed into abstract objects and (mental) properties: e.g., "belief". Physical objects are *per* Descartes (1996) primarily spatially extended: i.e., they can be described in an observer-independent way in terms of length, width, and height[8]. Of course, they can be located temporally as well. Alternatively, they can be described in an observer-dependent way in terms of e.g. the observer's distance from the object, the observer's physical orientation with respect to the object, and the observer's perspective on the object: what aspects the observer is focusing on, what aspects the observer is ignoring.

As Hemeren (2008) notes and Machery (2009) corroborates, much of the research into concepts, particularly in psychology, has focused on object concepts – generally concrete and not abstract ones – and their relationship to nouns in language. This may make them in some ways better understood than other sorts of concepts, but there is potentially a double distortion going on: a conflation of object concepts with nouns (and therefore with language), and a tendency to view all sorts of concepts through the lens of object concepts.

As Alex Martin suggests, object concepts are paradigmatically about physical objects, and physical objects are paradigmatically things we experience through our sensorimotor engagements with them. Physical objects have properties that are typically couched in directly (e.g., form, colour) or indirectly (e.g., weight) sensory-based terms. They exist in a context of motion and of motor descriptions: how the objects typically move (or do not move), how they interact (or do not interact) with their environment and with the agents observing them, how (in the case of tools) they may (and may not) be used. As Martin points out, the sensorimotor approach to object concepts is far older than the recent surge of interest in embodied cognition might suggest.

For more on the sensorimotor grounding of (all) concepts (not just object concepts), see Section 7.2.3.

[8] One might argue that there are concepts that intuitively belong in this category but lack these three dimensions, and on first appearances, one would be right. A point has neither length nor width nor height: zero dimensions. A line or line segment has length but neither width nor height: one dimension. A plane figure has length and width but no height: two dimensions. Yet a point, a line segment, and a plane figure are all abstract mathematical concepts. As they are encountered in perception – a point as e.g. the period at the end of the last sentence, a line as e.g. a line drawn with a pencil, a plane figure as e.g. a painting or a photograph or an image on a computer screen – they always have three dimensions, and must, in order to be perceived. It is only that one or more of the dimensions is conceptually irrelevant. For the photograph or the image on the computer screen, the depth or thickness is irrelevant. For the period at the end of a sentence, only its relative location is relevant.

4.2.2.2 Actions and Events

> Review articles... and books on concepts and categorization are remarkably silent about action concepts. The obvious question is: why is this so? One obvious reason is that there is little research done on action concepts. The follow-up question is then: why is there so little research on action concepts? Actions are difficult to study. They are dynamic and easily confounded with other variables (Hemeren, 2008, p. 48).

In keeping with the literature, Hemeren writes of *action concepts*. I will prefer instead to talk about action/event concepts as I find this expression more inclusive. Actions I mean in the conventional sense of being "intentional under some description" (Davidson, 1980, p. 50) (i.e., by an agent), as opposed to other sorts of events, which are "mere happenings": a man shouts, a volcano erupts. Davidson is the standard reference here, and I defer to his notion of agency. For my purposes I will not find it useful to distinguish between the two.

As with objects, action/events, toward one end of the continuum, will simply be physical action/events: i.e., action/events involving physical objects and locatable in physical space; toward the other end, again, they will bleed into abstract entities and (mental) properties: e.g., "believe". Action/event concepts are temporally extended: i.e., they can be described in an observer-independent way in terms of duration and likewise in terms of might-have-beens (what would have happened if one or another circumstance had been different). Of course, they can often be given a spatial location as well. At the same time, they can be described in an observer-dependent way relative to the observer's distance in space and time from the action/event and, once again, the observer's perspective on the action/event: what aspects the observer is focusing on or ignoring.

As they relate to physical objects, action/events have properties that are, as with physical objects, couched in straightforwardly sensory-based terms: visual (velocity, acceleration, bearing), auditory (volume, bearing), and so on. How fast did she throw the ball? How loud did he laugh? They exist in a context of the physical entities who engage in them or whom they are directed at or otherwise involve. As they relate to abstract entities, they have properties that are frequently (though not always) couched in metaphorically sensory-based terms: *fueling* the frenzy of the mob, *cultivating* hope. A man (the physical entity) throws a ball, but a man (the abstract mental entity) tells a lie. Of course, we treat them as being the same man: hence, my earlier point about blurring the distinction between mental and physical.

As Paul Hemeren notes, much if not most of the research in psychology has focused on object concepts to the negligence of action/event concepts. Consider, for example, the research on concept acquisition in early childhood (Johnson et al., 2003; Rakison and Lupyan, 2008), in adults (Schyns et al., 1998), and in non-human animals (Hall-Haro et al., 2006). As one of the reasons for this, he notes the dynamic nature of actions – ironic, given what I have suggested in Section 3.2.4 and will argue in Chapter Seven is the dynamic nature of concepts, for all their stable appearances. If I am right, then action/event concepts merely make these dynamics more explicit.

One might argue, in favour of the focus on object concepts, that object concepts and action/event concepts are entirely different sorts of things, with different properties and different structures – as it were, to borrow a turn of phrase from Jolley (see Section 2.5.1), "a distinction without a genus". I believe that Machery (2009) holds this view. Hemeren disagrees: "In order to develop an understanding about the organization of action categories, it may be the case that action categories

share a similar structure with object categories" (Hemeren, 2008, p. 25)[9]. This is, indeed, what Hemeren found with a series of well-designed experiments: action categories show a similar if somewhat simpler structure, similar if somewhat shallower hierarchical organization, analogous basic-level effects (see (Rosch, 1999)), similar typicality effects, and so on.

4.2.2.3 Properties

> In most semantic theories or theories of concept formation, no distinction is made between *properties* and *concepts*. I propose, however, that properties should be seen as only a special case of concepts (Gärdenfors, 2004, p. 60).
>
> ... A property is based on *one* domain (a subspace of integral dimensions[10]), while a concept may be based on *several* domains (consisting of separable domains) (Gärdenfors, 2004, p. 101).

Again, in keeping with the literature, Gärdenfors writes of properties as both the concepts of properties (i.e., a particular type of concepts) and the referents of those concepts, with the meaning disambiguated by context. I will prefer instead, where I wish to refer to the concepts *qua* concepts, to talk of "property concepts" whose referents are properties.

Unlike either objects or action/events, properties are, on the account I am offering, more or less strictly toward the "mental" end of the continuum, bleeding, at the lower end of the spectrum, into abstract objects and action/events[11]. *Contra* property dualism, properties are (all) abstract, mental entities. Remember the point that Barrett made earlier (Barrett, 1962, p. 130): concepts are (all) abstract, mental entities. So on this account, and as I am using the terminology, all concepts will be properties, but only some concepts will be property concepts. (This is not strictly true. I will argue in Section 6.2.2.3 that all concepts *can* – in certain circumstances – be treated as property concepts.)

If physical objects are spatially extended and action/events are temporally extended, then properties are conceptually extended, within a *conceptual space* that is the analogue of physical space. To borrow an example that Gärdenfors frequently uses (including many times in (2004)): in the case of color, the integral dimensions of hue, saturation, and brightness define a cone-shaped conceptual space (the familiar colour cone – see Figure 4.1).

Properties may themselves have properties. A shade of brown may be light or dark. A weight may be heavy or light. Likewise, a mind may be described in terms of its intelligence, a particular consciousness in terms of whether it possesses reflective self-consciousness, a concept in terms of whether it is first- or higher-order. So properties exist in a context of objects *and* action/events *and* properties, since they can attach to all three.

[9] I distinguish concepts from categories (see Section 3.4.3); roughly, categories are an aspect of concepts; categorization is one thing that concepts do. Hemeren makes a similar distinction, I think. Here, it is enough to note that if action/event categories and object categories are similarly structured, then *ceteris paribus* action concepts and object concepts will also be similarly structured.

[10] Integral dimensions are ones where you cannot have any one dimension without the others: e.g., in the case of colour (a property concept), you cannot have any one of hue, saturation, or brightness without the others.

[11] Abstract objects like mind or consciousness, or any of the propositional attitudes, can, in this way, be looked at as properties of physical organisms or of brains.

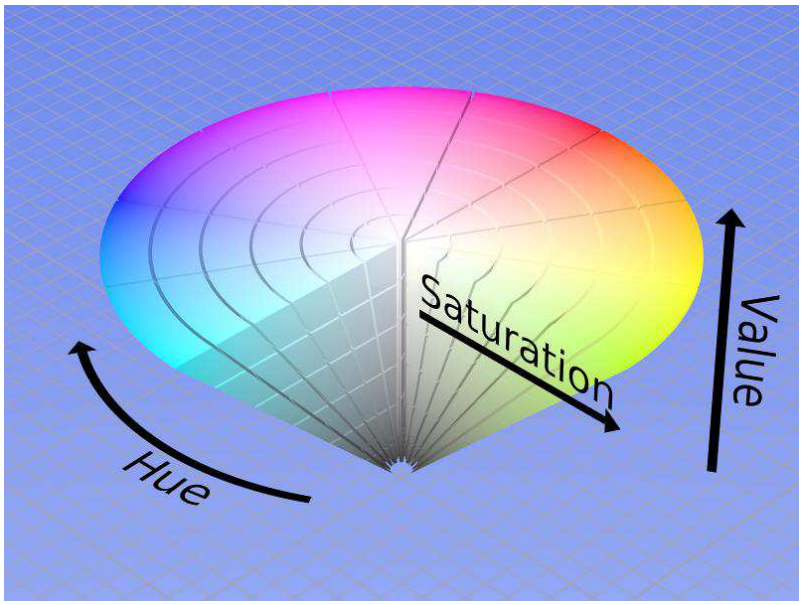

Figure 4.1: The colour cone. (Photo downloaded from Wikimedia Commons: http://commons.wikimedia.org/.)

4.2.3 Homogeneous versus Heterogeneous

I have one final distinction to make between categories of referents, which will likewise prove useful to the discussions in Chapter Six: to wit, between things that are homogeneously and things that are heterogeneously structured. Take water, which can be divided into smaller amounts of water, which are every bit as much watery as the water from which they are subdivided; or a ball of well-mixed cookie dough, which can be divided (up to some practical limit) into ever smaller balls of cookie dough. Many things, however, are not homogeneous. A tree cannot be divided into trees (unless perhaps it is a baobab tree) but instead consists of root, trunk, branches, leaves, and so on. Likewise, a bicycle cannot be divided into smaller bicycles but is instead an assembly of specialized parts: the wheels, the chain, the pedals, and so on.

(On the other hand, a bicycle *qua* artefact *is* homogeneously structured: it is an artefact composed of artefacts. As a simple point of logic, any heterogeneously structured entity *can* be viewed as homogeneously structured, given the right – i.e., sufficiently general – perspective. Likewise, any homogeneously structured entity (other than a singularity) *can* be viewed as heterogeneously structured, given the right – sufficiently specific – perspective. It all depends on what category one chooses as a contrast class. For a discussion of the contextualizing effects of contrast classes, see (Gärdenfors, 2004, pp. 119-122).)

This does not apply only to physical objects; it applies as well to abstract things like language, as noted in Section 2.4.1; or to concepts (or, if you prefer, conceptual mental content). On most if not all accounts, *some* conceptual content can be decomposed into smaller amounts of conceptual content (i.e., some concepts decompose into appropriately related "simpler" concepts) and some cannot. The same holds for action/events. An act of walking can be divided into smaller acts of walking: I walk across campus by first walking across the room. Jumping out of the way of an oncoming car cannot be divided this way. The jump consists, not of smaller jumps, but of a complex series of coordinated motor responses that together produce the jump.

Oddly enough, the same does *not* hold for properties, except at the boundaries where they bleed off into abstract objects or action/events. Properties are structured, of course; but they lack homogeneous structure. Objects can be structured out of other (self-similar) objects and action/events out of other (self-similar) action/events, but it is in the nature of properties that they are not, in any way, structured out of properties: a property of brown or heavy or sweet *just is* a property of brown or heavy or sweet. More will be said in Section 6.2.2.3.

The distinction between homogeneously structured things and heterogeneously structured ones is very close to the one between classes (or *categories*) and instances. For a category to be a category in any useful sense, its members must (at some appropriate level of abstraction) be homogeneous: i.e., they should have a set of essential properties in common[12]. So for example, however one divides up the category of mammals, by whatever criteria, one should end up with (smaller) groups of mammals. This is precisely what I am claiming, in this work, for the category of concepts writ large[13] – in contrast to someone like Edouard Machery, for whom "the class of concepts divides into kinds that have little in common" (2009, p. 5). For more on the class/instance distinction, see sections 2.1 and 6.1.

Of course, any homogeneous structure must, on pain of infinite regress, at some point "bottom out". In practical terms, a quantity of water cannot be divided beyond the point of individual drops. Even if it could be, at some point one would reach the level of individual water molecules, which consist not of further water molecules but of hydrogen and oxygen atoms. The "primitive element" for water is either a single droplet or a single water molecule, depending on how it is being considered. The "primitive element" for walking is a single step. In either case (and perhaps in any case), the "primitive elements" can be treated as unstructured (which, relative to the original domain of discourse, they are), *or they can be treated as structured relative to a different domain.* The water droplet can be analyzed into molecules, the molecules into atoms. The step can be analyzed, like the jump, into a series of motor responses.

4.3 Types of Concepts by How They Are Used

> The prime problem is that the information received by the receptors is too rich and too unstructured (Gärdenfors, 2004, p. 221).
>
> In other words, there could be two different, but equally real, patterns discernible in the noisy world. The rival theories would not even agree on which parts of the world were pattern and which were noise, and yet nothing deeper would settle the issue (Dennett, 1991b, p. 49).

All of the preceding discussion may make concepts sound static and not, as I argued in Section 3.2.4, dynamic entities; to recall the discussion from Section 2.7: more like representational entities to be described and less like abilities to be performed. But of course, as I argued there, concepts are as much abilities as they are representations. So something needs to be said not just about *who* is possessing and employing the concepts, and *what* they are being attached to, but also *how* they are being used.

[12] A folk category widely understood in academia to be a heterogeneous and therefore erroneous category is race. Despite presumptions otherwise, the category has no discernible biological basis.

[13] . . . As does probably any philosopher who sees any point in putting forth a theory of concepts: see e.g. (Machery, 2009, p. 18).

As I will argue at length in Chapter Seven, concepts structure the very experience that gives rise to them. I wish only briefly to sketch the idea here. *Per* Gärdenfors' suggestion, concepts take a sensorily rich but unstructured cognitive level of basic sensorimotor engagements between agent and environment and give it structure, finding patterns in the raw percepts and then patterns in those patterns, dispensing with the irrelevant or less relevant, clumping together that which remain into things that eventually look like e.g. objects, actions, and properties: the equivalent of, in computer graphics terms, a *lossy compression* algorithm[14]. This is what, according to Dennett, makes a pattern a "real" pattern (and nothing more!): its ability to achieve data compression.

Concepts allow the conceptual agent to step back from strict experience-in-the-moment, to gain predictive advantage by finding patterns across many moments and many contexts – an idea with its roots in e.g. (Goldstein and Scheerer, 1941) and echoed in many places since. The proof of a good ("real") pattern is in its capacity to predict; of a good concept, its applicability across many contexts.

(One could deny, of course, that concepts relate to experience in this sort of way, in keeping with the rationalist as opposed to empiricist tradition. Certainly Fodor is inclined toward this view. It should be clear by this point that my sympathies are broadly with the empiricists. My purpose in this work is not to reject rationalism. Rather, it is to provide some positive account of why, Fodor's claims to the contrary (e.g., (2008, p. 28)), the "empiricist project" is *not* dead.)

Of course, if concepts structure the experience that structures them, they may do so in different ways, playing different roles or taking on different appearances depending on how they are doing so. At different points through the course of this work, I will talk about concepts in all three of these ways[15]:

- **components:** Some things are part of something else, themselves composed of parts, or (most commonly) both. Sometimes we look at concepts that way as well, in which case we might think of them as *building blocks* (*cf.* (Rumelhart, 1980)) to be piled up, torn down, and piled up again.
- **parameters:** Some things are features or aspects or parameters of other things: they say of what they describe or model that it is like *this*. When we look at concepts this way, they are *properties* (see Section 4.2.2.3) of things.
- **contextuals:** Some things are part of the background for other things that are in the foreground: not themselves there to be described, but just "part of the scenery". Seen in this light, concepts form the context that we, as ourselves conceptual agents, cannot step outside of; they are the ever-present *backdrops* to our experience of the world.

I will make use of this tripartite division in a very different way in Section 6.2.2, as a means of formally describing the structure of individual concepts.

[14] A lossy, as opposed to loss-less, compression algorithm is one that makes a trade-off between level of compression on the one hand and recoverability of some semblance to the original data on the other.

[15] I first suggested such a division in (Parthemore and Taylor, 1992).

4.4 The Evolution of Concepts

> No truth appears to me more evident, than that beasts are endow'd with thought and reason as well as men (Hume, 2003, p. 126).
>
> Humans are undoubtedly unique in their spontaneous invention of language and symbols; but, as I have argued elsewhere... our special advantage is more on the production side than on the conceptual side of the ledger. Animals know much more than they can express (Donald, 1998b, p. 185).

There is quite a different sort of context altogether in which we can locate concepts – one that will reinforce my argument (from Section 3.3.2) that concepts and language pull apart and help us understand better the continuities between human and non-human cognition. This is, of course, an evolutionary context.

It is as unwise as it is common to talk about the nature of concepts in philosophy of mind without considering both the advent and the evolution of concepts: i.e., the genesis of conceptually structured thought from its nonconceptual origins, and its development from that point forward. When these matters are considered at all, it is usually in the context of the individual conceptual agent and not the species. Yet unless one is prepared to grant conceptual abilities to even the simplest organisms, at some point conceptual genesis must have taken place; and unless one believes that all conceptual abilities are equivalent, some account must be made of the way conceptual abilities, if not the concepts themselves, change.

Donald's work on the evolution of cognition (1993; 2001b) provides an excellent foundation on which to ground such a discussion while providing support for the positions I myself want to take. He is keen to stress the continuity between non-human and human cognition, the better to highlight those aspects of human cognition that *are* distinctive. His four stages of "cognitive-cultural development" – episodic, mimetic, mythic, and theoretic – show how a conceptual foundation common to many species is progressively transformed and becomes, in humans, itself a means toward cognitive evolution.

4.4.1 Advent of Concepts: A Baseline

I have already laid out, in sections 3.3.2 and 4.1, my thoughts for a baseline above which agents can appropriately be considered conceptual agents. Although Donald writes within a very different terminological sphere from the literature on concepts, nonetheless I think it clear that, on these criteria, some and possibly many individuals and species qualify as conceptual agents who have not passed through any of his four stages, which he wishes to limit more or less (and I think correctly) to the higher social mammals.

4.4.2 Conceptual Transformations

Those stages are about not just the progressive transformation of cognition (specifically, *per* my interests, conceptual cognition), but at the same time the progressive emergence of culture out of the most basic social elements. So the cognitive transformations – at first strictly genetic, in the end strictly social/cultural – are from the beginning dependent on a social context and would not arise (or at the very least, would take an entirely different route) in a non-social species. They are *cultural* transformations. The key features of this model are:

- The conservation of previous gains. Each stage builds upon rather than replaces the last.
- The greater stability of older over newer cognitive systems.
- The important position of mimesis as the oldest of the uniquely human adaptations.

My interest here is in how conceptual abilities are progressively transformed, and how new concepts and new classes of concepts arise that could not have done so before.

4.4.2.1 Episodic Culture

> Episodic memory is, as the name implies, memory for specific episodes in life, that is, events with a specific time-space locus. Thus, we can remember the specifics of an experience: the place, the weather, the colors and smells, the voices of the past.... Such memories are rich in specific perceptual content. By definition, episodes are bound in time and space to specific dates and places (Donald, 1993, p. 150).

The social mammals show significant conceptual advances over other agents we might consider conceptual agents, such as the parrot Alex, in several key areas. The great apes in particular are able to:

- While not creating tools of their own, show great readiness to take advantage of tools that they find.
- Invent solutions to problems, such as how to get at some food.
- Solve so-called delayed reaction tasks, where the agent must wait for a reward.
- Recognize themselves in a mirror (Donald, 1993, pp. 124-126).

All of this suggests a move from more passive toward more active, more intentional cognition. What ties all of these traits together is, Donald argues, significant advances in *event perception* and the advent of *episodic memory*. Together these make possible episodic cognition and *episodic foresight.*

Event perception "is, broadly speaking, the ability to perceive complex, usually moving, clusters and patterns of stimuli as a unit" (Donald, 1993, p. 153). Event perception can be placed along a continuum: "Animals that we call intelligent are those that respond to events of increasing complexity and abstraction. Apes can discriminate hand signs that are too complex or subtle for dogs; but dogs can read aspects of behavior that are missed completely by rats"(Donald, 1993, p. 153).

Episodic memory (the term dates to (Tulving, 1972)) ties objects and events together. It has an interesting dual nature: on the one hand, it is, as the quote from Donald emphasizes, highly specific to a particular occasion; on the other, it allows agents who possess it to relate the objects and events comprising an episode to each other in ways that agents without episodic memory cannot. Donald places the ability to conceptualize about action/events as being higher order than conceptualizing about (physical) objects. Episodic memory carries on this progression, allowing objects not only to be successfully re-recognized but also associated with each other through various action/events, and for actions/events to be associated both with each other and all the objects they involve as parts of unified episodes.

Episodic foresight goes a step further: "experiencing a mental episode that is acted upon as if it were a possible future for the subject", a kind of "mental time travel" (Osvath, 2010, p. 9). Mathias Osvath describes a series of experiments conclusively showing the existence of episodic foresight in enculturated apes. In 2009, Osvath and his team made international headlines for a case study of quite elaborate planning behaviour by a male chimpanzee at Furuvik Zoo near Stockholm (Osvath, 2009). Episodic foresight comes close to, if not in fact requiring, some of the same intentionally reflective capacities as Donald reserves for mimetic culture.

Episodic foresight aside for the moment, I believe episodic culture, as Donald describes it, offers the first unambiguous signs of implicit – not yet explicit – meta-cognitive abilities: i.e., the first signs of thoughts about thoughts (higher-order thoughts) without need for any awareness by the agent of such thoughts. Episodic cognition draws a line between recognizing e.g. a tree *as* a tree (insofar as being able to re-identify it reliably) and recognizing it *as* a tree *in the context* of a wider setting that involves other plants, other agents, various activities involving that tree, and so on.

It is with the advent of episodic cognition that the first recognizable animal *cultures* emerge, or what Donald refers to as *social intelligence.* The cultures of the higher mammals are distinguished from so-called social insects by the flexibility that is the hallmark of conceptual abilities. On the one hand, they could only arise once those basic conceptual abilities were in place; on the other, they make possible the progressive transformation of those abilities far beyond their starting point, in part by blurring the lines of where one agent's concepts stop and another's begin, or between where one agent's concepts stop and those of the social group begin.

4.4.2.2 Mimetic Culture

> Cognition is traditionally identified at the level of single individuals – this might be termed the assumption of the "isolated mind" – and in other species, this assumption seems largely justified, since non-human species do not have a capacity for intentional representation, and are thus unable to transmit acquired knowledge across generations (Donald, 1998a, p. 11).
>
> If I see somebody use a stone as a tool to crack open the shell of a nut, I may do the same thing, not to bring into mind the act of the other person I have observed, but to obtain the same effect (Sonesson, 2009, p. 12).

Mimesis – paradigmatically characterized by gesture – begins the social transformation of concepts, which oral and written language then carry much further. That is to say, it takes place *in a social context*, where agents are actively a part of each other's conceptual learning process. It is, on Donald's account, the first of the strictly human stages. Donald distinguishes mimesis from imitation or mimicry by its intentionality and its representational nature. As Göran Sonesson notes, imitation is necessary but not sufficient for mimesis.

Key to intentional representation is *explicit* meta-cognition, which Donald offers as a necessary precondition for mimesis (and which may, if Osvath is right, already be present at the earlier stage). This is directed reflective thought, enabling the agent to go beyond merely abstracting on pre-existing concepts into wider and wider contexts. Remember my earlier point (Section 2.6) that it is only where there is conscious intent that one should speak of representation *from the viewpoint of the agent* (as opposed to e.g. the external observer) at all. Representation is an *active* process requiring an aware observer.

With mimesis the representations are largely, if not entirely, *iconic* and not yet *symbolic representations*. Iconic representations retain a discernible link back to their sensorimotor origins and so are easy to reproduce and communicate by sensorimotor re-engagement.

Mimesis makes possible the *sharing* of concepts. In so doing it creates, for the first time, a distinction between private and public aspects of concepts (see Section 3.3.3) – with a non-trivial mapping from one aspect to the other. Such sharing can then have profound effects back on the cognitive abilities of individual agents, allowing them to understand without having directly to experience, as well as constraining what they do experience (see e.g. (Gärdenfors, 2004, p. 190)).

Unlike episodic culture, there are no extant examples of mimetic culture, making it impossible to prove even whether it has ever existed. Yet it very conveniently fills a gap in the evolutionary story, so that abstract concepts and language abilities need not appear out of nowhere. One can speculate that here, for the first time, one finds:

- Rehearsing and modeling of society, where children can act out not only their own roles but those of others in their society.
- Structured games, with rules.
- The emergence of ritual, including dance.
- Complex acculturation of the young resulting in pedagogy (Donald, 1993, pp. 174-176).

4.4.2.3 Mythic Culture

> The mind has expanded its reach beyond the episodic perception of events, beyond the mimetic reconstruction of episodes, to a comprehensive modeling of the entire human universe. Causal explanation, prediction, control – myth constitutes an attempt at all three, and every aspect of life is permeated by myth (Donald, 1993, p. 214).
>
> . . . Language has had a privileged place in human culture and human thought, as shown by the fact that the name of a given language and that of the people speaking it are nearly always the same (Zlatev, 2009, p. 187).

"Mythic" is the name Donald gives to oral-language-based cultures which, unlike mimetic cultures, have survived into modern times. The affect of language on social and individual cognition is difficult to overstate: "Simultaneously with the appearance of speech there appeared a whole constellation of thought skills that are associated with language and are, broadly speaking, linear, analytic, rule-governed, and segmented" (Donald, 1993, p. 212). At the same time, it would be a mistake, on Donald's account, to see language as leading the cognitive changes. Rather, ". . . symbolic thought is *primary*; it is the driving force, the invisible engine, behind word use" (Donald, 1993, p. 233). Symbols do not make structured thought possible, but, with their seeming arbitrariness, they do represent a radical shift in representational strategy.

It is worth remembering that humans are not alone in the ability to learn symbols, understand the arbitrary relationship of sign to signified, and therefore be able to employ symbols appropriately. Such ability has been shown quite clearly in studies with the social mammals. What seems unique to humans is the ability to invent symbols spontaneously – something that, according to Donald, requires a mimetic foundation. "The most likely initial source of arbitrary symbols in mimetic

culture would have been in the standardization of mimetic performance – that is, in gesture" (Donald, 1993, p. 220).

In mythic culture, symbolic representation in general and language in particular serve a primarily *integrative* function. "The most elevated use of language in tribal societies is in the area of mythic invention – in the construction of conceptual 'models' of the human universe" (Donald, 1993, p. 213). Again: "... although language was first and foremost a social device, its initial utility was not so much in enabling a new level of collective technology or social organization, which it eventually did, or in transmitting skill, or in achieving larger political organizations, which it eventually did.... Its function was evidently tied to the development of integrative thought – to the grand unifying synthesis of formerly disconnected, time-bound snippets of information" (1993, p. 215). So the integrative role of concepts, which began with episodic culture, tying together the different components and aspects of an episode, reaches its apex, perhaps, in mythic culture, where it is the entire world and one's place in it that are being brought together.

At the same time and in contrast, with each stage in cognitive-cultural development, the concepts themselves – or rather, their most visible aspects – become increasingly structurally impoverished, *even as they become more visible to the agent*. After all, representations of any kind are generally, if not uniformly, simplifications of their representeds; and symbols, in the limit, take that simplification to an extreme.

4.4.2.4 Theoretic Culture

> This, again, would seem to be a breaking point on the way to human beings: the possibility of memory as an external record, which perdures independently of the human organism (Sonesson, 2009, p. 3).
>
> Writing is really a way of transferring the storage of an idea from the brain (its natural resting place) to a non-biological medium. Ideas started in the brain, where they traditionally resided through most of human history (Donald, 2001a, p. 559).

The last of Donald's four stages, and the last of the three distinctively human ones, is also the *first* that is not genetically but strictly culturally mediated, arising not spontaneously (as oral language appears able to do) but only with the appropriate pedagogical enculturation. It is also the first to introduce, not new conceptual machinery, but what would seem (if the account offered earlier is correct) a conceptual fiction: the idea of conceptual knowledge as detached from any particular agents and perhaps *from any agents at all*. Mythic culture is unabashedly subjective, its conceptual model of the world straightforwardly anthropocentric. Theoretic culture lays claim to "true" objectivity, elements of human perspective corrected for and removed. Mythic culture is occupied with telling stories, and aims for the "big picture"; theoretic culture is occupied with revealing logical truths, and concentrates on the details. If mimesis made possible the sharing of concepts and oral language extended that capacity, then written language takes concept sharing to a point where one might forget that concepts are anything *other* than shared, public entities.

Besides the externalization of memory – which Donald means quite literally, in the spirit of Andy Clark and David Chalmers' "extended mind" hypothesis (Clark and Chalmers, 1998; Clark, 2008) – Donald sees two other significant cognitive deficits in oral-mythic culture: *graphic invention* (the creation of visual images with symbolic intent) and *theory construction* (the development of

carefully constructed arguments based on logical analysis and empirical discovery). "The major products of analytic thought... are generally absent from purely mythic cultures. A partial list of features that are absent include: formal arguments, systematic taxonomies, induction, deduction, verification, differentiation, quantification, idealization, and formal methods of measurement" (Donald, 1993, p. 273).

All three deficits are, of course, closely related. Although early cave art is quite different from writing and proceeded it by thousands of years, nonetheless they are both graphic inventions that create an external record capable of surviving far beyond the lifespan of any individual, one that permits verification in a way that oral narrative does not.

The analogy Donald makes for the increasing reliance of the individual agent's cognitive abilities on those of the group is between a standalone and a networked computer: unlike the standalone computer, the specifications of the networked computer (in terms of random-access memory, hard drive capacity, and so on) may not tell you much. By "plugging into, and becoming a part of, an external symbolic system" (Donald, 1993, p. 274) agents can offload many cognitive tasks, particularly those involving memory, onto external resources. "The mnemonic arts and rote learning, once a major part of education... have receded into the background as the reliance on biological memory for storage has faded" (Donald, 1993, p. 323).

The free marketplace of permanently recorded concepts has profound effects back on the private conceptual life of the individual. One can acquire many new concepts merely by reading about them. Those concepts that are acquired by direct personal experience can be further shaped, and re-shaped, through the new media.

A more subtle effect on the individual is the cultivation of increasing levels of self-reflection. "The shift was away from immediate, pragmatic problem solving and reasoning, toward the application of these skills to the permanent symbolic representations contained in external memory sources" (Donald, 1993, p. 335). So there is a further stepping back from the particulars of the moment, from perhaps even the possibility of practical application, to respect and appreciation for "reflection for its own sake" (Donald, 1993, p. 341).

Stage	**Development**
episodic	implicit meta-cognition appears
	concepts take on integrative role
mimetic	explicit meta-cognition appears
	iconic representations appear
	private/public distinction appears
mythic	symbolic representations appear
	oral language appears
	integrative role for concepts reaches its apex
theoretic	concepts become externally "free floating"
	written language appears
	public aspect of concepts dominates

Table 4.1: Stages of cognitive-cultural development and corresponding conceptual innovations.

It will be useful at this point to summarize, in Table 4.1, what conceptual changes I have derived from Donald's account for each of his four stages.

4.4.3 The Difficulties of Looking Backward

It might be tempting to see each stage of cognitive-cultural development as an improvement over the last, but a more sober analysis suggests a succession of trade-offs. If it is indeed in the nature of *all* concepts and all conceptual abilities that they permit the agent to step back from strict experience in the moment, to consider the present moment in light of past or future moments – as I suggested in the introduction – then the first trade-off with concepts is flexibility of response in exchange for a loss (literally) of sheer impulsiveness. It is the loss of spontaneity in the precise opposite of Kant's sense. If this is right, then Donald's statement that apes' "lives are lived entirely in the present" (1993, p. 149) should be understood with the caveat that that "present" has already been stretched far beyond its pre-conceptual boundaries. Rousseau (2004) was neither the first nor the only writer to suggest that our lives might have been better off if we had held onto that un-Kantian spontaneity. But as Rousseau himself acknowledged, having moved on, we cannot go back. The alienation from the present moment only becomes more pronounced with each further stage of development.

Representations likewise suggest a trade-off. The increasing role of representations through mimetic, mythic, and theoretic culture point to their power. At the same time, representational cognition is costly, and it is slow.

Iconic representations are meant to evoke, directly, a certain sensorimotor association in the recipient, a sense of resemblance to their representeds. With symbolic representations, that link back to sensorimotor-grounded origins is lost, the form apparently arbitrary in relation to its function. Their simplified structure both makes them extremely efficient in terms of storage space or amenability to rule-based processing and at the same time critically dependent, far more than iconic representations, on a shared context for understanding, a common conceptual space.

Finally, the new conceptual apparatus provided by theoretic culture has been fantastically effective at systematizing and preserving knowledge beyond the lifetime of the individual or even the society. At the same time, it is founded on the fiction that conceptual knowledge can be made independent of the agents possessing and employing it and of the influence of their perspective. However useful that fiction is, it is not one that should be taken as atemporal fact. Harvey's comment comes to mind: "the underlying assumption of many is that a real world exists independently of any observer; and that symbols are entities that can 'stand for' objects in this real world in some abstract and absolute sense" (1992, p. 5).

By being so successful with the details, theoretic concepts have, perhaps, lost sight of the "big picture". Individual agents become increasingly specialized within the corporate conceptual structure. Who, even among climate scientists, has the "big picture" on human-mediated climate change? Who, if anyone, is qualified to assemble all the pieces that would constitute a sustainable relationship with the environment? Then, too, not all human knowledge fits into an analytic mold. It can be easy to ignore, or downplay, or disparage, that which does not fit the requirements of logical argument or empirical discovery. In modern society, "myth" has become a pejorative.

There may be a simple reason why it is easy to perceive the gains and difficult to see the trade-offs. With each stage of cognitive development we pass through – individually or as a species – it becomes at the least difficult and in the limit impossible to imagine what cognition was like previously. Once we have written language, it is difficult to imagine being illiterate; once we have

language of any kind, it is difficult (for many at least) to imagine thoughts without words; it is as impossible to take a non-representational view on pre-representational thought as it is to imagine structured thought outside the confines of episodic memory; and so on. The more we become comfortable with each new conceptual tool, the more it comes to feel like an essential part of us and not an extension at all: we integrate it into our core self-image.

4.5 Theories in Use

Before closing this chapter, it will be useful to put not just concepts but theories of concepts into some useful context. If concepts exist in a context of agents and referents, of evolution and use, then theories of concepts arise in some analogous context, most notably that of the applications for which they are intended. Although I do not share his pessimism about theories of concepts, on these points Machery (2009) is unquestionably right: researchers in different fields (and often enough, within the same field!) cannot simply assume that they are talking about the same thing when they talk about concepts; and a theory of concepts suitable for every research field and every application is probably impossible. In any case, it is not desirable. As the proverb goes, "a tool that does everything does nothing well".

Nonetheless, I think a theory of concepts that sits between psychological and philosophical accounts is a worthy and achievable goal. Cognitive science, which provides the backdrop for the theory of concepts I put forward in this book, has a long tradition of drawing on both psychological and philosophical theory, even as it seeks to put those theories to the test in practical, computer-based applications.

In the first chapter I made the claim that every research program in cognitive science and every AI project takes place in the context of some theory of concepts, whether that theory is stated explicitly or not (and most often it is not). In the 1980s the buzz phrase was "knowledge representation", which focused (perhaps far too singlemindedly in retrospect) on explicit and symbolically encodable knowledge. Although the phrase has fallen somewhat out of fashion, the interest in the research question, certainly within the philosophical side of cognitive science, remains strong, regardless of comments like this one from Machery:

> Once at the center of philosophy, the philosophy of concepts has now been marginalized, maybe because for a few years now, it has been stalled (2009, p. 3).

As evidence I cite the continuing excitement over the friendly competition between Jesse Prinz's proxytypes theory and Jerry Fodor's informational atomism. A major international conference on the philosophy of concepts was held in Copenhagen in May 2007, with all the leading researchers in the field in attendance, including Fodor, Prinz, and Gärdenfors; and a recent *Journal of Consciousness Studies* special issue (September-October 2007) was devoted to the interplay between concepts and consciousness.

It will be helpful at this point to consider two specific cognitive science research projects into the nature of conceptual knowledge. One is Doug Lenat's Cyc, a large-scale project which has been ongoing since 1984 (Lenat and Guha, 1989); the other is a small-scale project I myself was involved with at the University of Manchester in the early 1990s. Each offers cautionary lessons.

4.5.1 Cyc

> As Cyc is the first project of its magnitude, its designers have not been able to follow blueprints established by anyone else. Instead, they have made their own decisions about what categories of informations should be included, how this information should be encoded and indexed, what inference mechanisms should be provided, how consistency should be maintained, and more. Our major concern with the project is the way in which these choices were made.... It appears that in general the choices made by the designers of Cyc were "non-decisions", that they felt it was important to build something – anything – to see if it worked, than to contemplate at length what choices were in principle best (Elkin and Greiner, 1993, p. 43).

The approach of the Cyc project[16] to the nature of knowledge in general and concepts in particular might be summarized like this:

1. What the Cyc Project calls "common sense" knowledge is mainly conceptual. Furthermore, "common sense" knowledge can be programmed. Consider Marvin Minsky, a supporter of Cyc, quoted on the project website: "People have silly reasons why computers don't really think. The answer is we haven't programmed them right; they just don't have much common sense."

2. All conceptual knowledge can be propositionally expressed. A concept, to Cyc, is a discrete, explicitly represented, recursively defined sub-propositional structure, of which there are currently 300,000 in the Cyc database. Concepts have definitions of a sort, but they are flexible, being constantly updated in light of new information.

3. All propositional knowledge can be captured in a form of higher-order predicate logic.

4. Knowledge does not require an intentional agent.

Non-conceptualists, enactivists, and many others will have a problem with (1), while (2) will be anathema to anyone of the concepts-as-abilities mindset (see Section 2.5). Even more controversial is (3), given that higher-order logics are generally understood to be less well-behaved than first-order predicate logic, while Dale Jacquette has convincingly argued that standard attempts by higher-order logics to escape Grelling's Paradox fail (2005) (though for a counterargument see (Ketland, 2005)). I have effectively rejected (4) in Section 3.2.1.

But this immediately raises a problem, as Machery makes us aware, for the Cyc people and I may well be talking at cross purposes. After all, much of the time the stated goals of Cyc are improved data mining, better search engines, and data consistency checking, all of which it seems (albeit anecdotally) to do well. Cyc's notion of concept allows it, in certain contexts, the ability to give a good semblance of common sense propositional reasoning. Looked at this way, the Cyc project isn't concerned with what a concept *really* is (any more than I, as a pragmatist, am) but what it can do with the notion of concept it has. Claims like the following, from John De Oliveira, the president of the Cyc Foundation (also taken from the website), may perhaps be forgiven as so much marketing hype: "In the Cyclify project... we ask you to imagine a world in which every single person is given free access to *programs that reason with* the sum of all human knowledge".

[16] Their web page be found at http://www.cyc.com.

At the same time, Lenat has stated that once a critical threshold of concepts and propositions has been reached, the Cyc database will start reasoning with the full depth of human common-sense reasoning; and he has made recent public statements (e.g. (Lenat, 2006)) that the threshold is close. When the focus shifts from what the Cyc system *can* do to what it *will* do, from a system that mines data to one that is self-aware and that reasons, then its notion of concept may bear much closer scrutiny and be subject to understandable criticism.

4.5.2 The Pharos Project

What if the goal is not to re-create human conceptual abilities but to complement them? What if, for example, the goal is to allow users of a software tool to build external models of portions of their conceptual domains, the better to make areas of implicit knowledge explicit and to check knowledge for consistency?

That was the aim of the Pharos Project at the University of Manchester in the early 1990s, which consisted of two senior researchers and two research assistants, including myself (Parthemore and Taylor, 1992). Officially Pharos fell under the rubric of what was then called computer-aided personal interviewing (CAPI); specifically, we were seeking to develop tools to support the social survey design process. Dr. Peter Halfpenny, the sociologist on the project, estimated that designers would typically spend an average of at least an hour on the wording of each question and still discover problems in the completed survey[17]. It was felt that, if the designers could be assisted through creating an external model of their research domain, that this would then provide the basis for an "intelligent" writing environment that would help them spot problems in the wording and the overall construction of the survey much more quickly[18].

A *concept* was described as a sub-propositional entity that could either be viewed discretely or as a context-sensitive part of a larger, highly interconnected network; either as an indivisible atom or as itself encompassing an area of the larger network. It was meant to have a uniform specification no matter what sort of concepts one might want to represent: one standardized building block.

The aim being not autonomous artificial intelligence but collaborative AI-based support, we meant our ambitions to be modest. In retrospect, our goals were laughably anything but modest, not least given our limited time and resources. It was a lesson to me to be careful what I claimed. Nonetheless, some of the goals behind that project and my earlier MSc thesis inform the present work.

4.6 Conclusions

This chapter has attempted to put concepts in context by classifying them according to the conceptual abilities of the agents who possess and employ them, the referents to which they attach, and the uses to which they are put. Concepts removed from any of these contexts are probably incoherent. It has attempted to put conceptual abilities into an evolutionary perspective as well, the better to understand how concepts and language pull apart, and to understand how human cognition both is and is not continuous with non-human cognition.

[17] Personal communication.

[18] "Equipped with a model of the research domain and an extensive model of syntax, the writing environment will, we hope, provide meaningful advice and feedback on the structure and content of questions written by the designer" (Parthemore and Taylor, 1992, p. 90).

All conceptual agents are capable of entertaining what I have called, in keeping with the literature, first-order concepts: concepts of non-concepts. Agents with a certain level of reflective capacities are capable of entertaining higher-order concepts, including *n*th-order concepts, highly abstract concepts, and at least some versions of the **SELF** concept. However, on pain of paradox or banishment of self-reference, no absolute line can be drawn between first- and second-order or between first- and higher-order concepts. This reveals a fundamental limitation in the nature of conceptual thought: as conceptual agents, we can never stop being conceptual agents, never step aside from that role in order fully to understand our nature as conceptual agents.

I have taken a particular position on the so-called mind/body problem: mental and physical are neither two kinds of basic substance nor two kinds of properties. They are not in opposition at all, though there is an unavoidable tension between them. If they are two kinds of *anything*, it is two perspectives: two perspectives at either end of the same continuum that marks out non-concepts from first- and higher-order concepts. Dividing concepts into object concepts, action/event concepts, and property concepts, both object concepts and action/event concepts range from physical to mental along that continuum, whilst property concepts are highly abstract and, therefore, predominantly mental entities. My approach borrows elements from Papineau's conceptual dualism, from property dualism, and from physical monism, though it might best be described as a form of neutral monism. As physical object concepts are spatially extended, so action concepts are temporally extended, and property concepts are conceptually extended. This provides many of the necessary elements and sets the stage for an extended discussion of Gärdenfors' conceptual spaces theory and a set of proposals for extending that theory in Chapter Six.

Just as concepts exist in a context of agents, reference, and use, from which they cannot be removed, so, too, theories of concepts can only be understood in a context of a particular research domain and intended application. Two cautionary tales are provided. The moral of the story is to aim for the broadest workable theory, *but no broader*, and to be careful what you claim.

Chapter 5

The Limits of Concepts and Conceptual Abilities

> . . . Because of constitutional limitations of our mind, we shall never be able to achieve more than an explanation of the principle on which mind operates, and shall never succeed in fully explaining any particular mental act (Hayek, 1999, p. 34).

In Chapter Two, I talked about concepts as abilities (e.g., the ability to represent) and concepts as representations, and concluded that whether they are one or the other depends on perspective. When we reflect on our concepts, we see representations: that is the role they play in our reflections. Yet most of the time, it seems, we are not reflecting on our concepts at all, but merely getting on with employing them, in which case they seem best described as abilities, for who, remembering the quote from Harvey, would be doing the representing, and to whom? So, it seems, there is a critical distinction to be made between concepts as we reflect upon them and concepts as we possess and employ them non-reflectively: not two distinct entities, but one entity with two distinct perspectives.

In Chapter Three I asked whether concepts *must* be able to be reflected upon by the agent possessing them and whether they must be articulable by that agent. I concluded that, within the picture I am developing, they are not. I concluded further that, in the human case at least, most concepts have both a private and a public aspect: the private aspect shaped more by our personal experiences, the public aspect shaped more by the collective experiences of our society and species. The public aspect relates intimately to our lexicalized concepts, the private aspect only indirectly. Yet it seems essential to our conceptualizing that, most of the time, we conflate the two, taking e.g. our **DOG** concept to be the next person's **DOG** concept, and both to be the concept lexicalized by the word "dog".

In Chapter Four I confirmed a critical distinction between first- and higher-order concepts, arguing that they can be distinguished by the level of reflective capacities required. I concluded that, once one allows higher-order concepts, an inevitable ambiguity arises about where, precisely, on the scale to place any particular concept, in part because a first-order concept, reflected upon, *is* by the process of reflection higher-order. Attempting to draw a sharp line between first- and higher-order, or to set out clearly defined categories of first-, second-, third-, etc. order requires either banning self-reference or inviting paradox. The problem, I concluded, is not the fault of the concepts

themselves, but of our inability to step outside of our role of observer – to those concepts or to anything else.

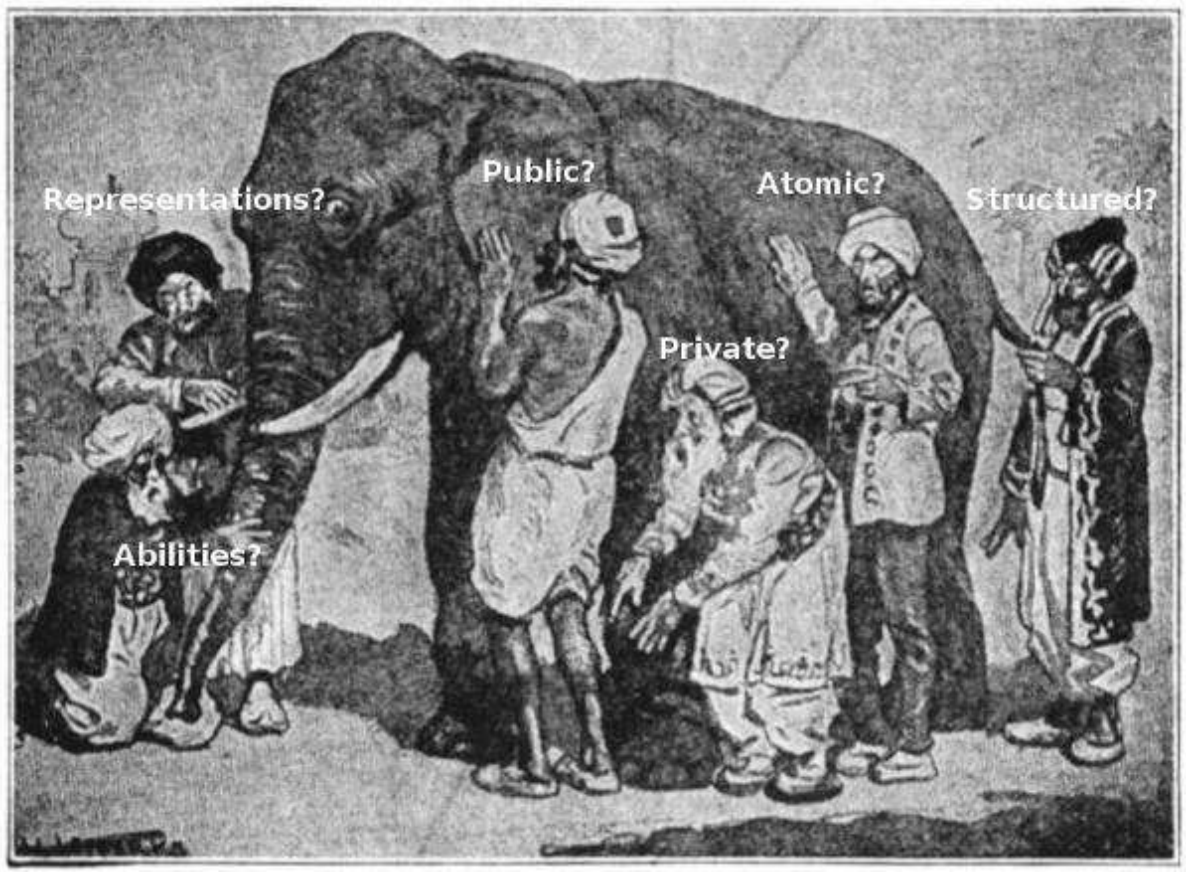

Figure 5.1: The Conceptually Blind Men and the Elephant. (Photo downloaded from Wikimedia Commons: http://commons.wikimedia.org/ and edited.)

This chapter will attempt to draw these strands together and argue that a too-complete account of concepts, one that attempts to account for everything fully, invites inconsistency – but that, in the end, relative completeness may be more important than strict consistency. As with the Sufi story of the blind men and the elephant (see Figure 5.1), conflicting accounts need not always mean either that those involved are talking about different things, or that one is right and the other(s) wrong. Different accounts may be differences of (limited) perspective and not differences of substance.

The blind men are all discussing the same thing – an elephant; and, as it happens, their accounts are all equally right – and equally incomplete (and in that sense, wrong). Like us, they lack the ability to take in the whole picture: they because they are visually blind, we because of our conceptual "blindness": our inability, even for a moment, to set aside our conceptual nature. But an inability to establish a final answer unbiased by perspective need not translate to unbridled relativism; some answers will be logically consistent with themselves, explanatorily useful, and match the available evidence; others will not. No one thinks that a concept is a static definition anymore and with good reason; as Fodor dryly points out, no one can find one (1998, p. 92).

If the first major contribution of this work is a pragmatic and distinctive account of concepts in terms of their essential nature, properties, and context of application, then the second is this: that concepts *by their nature* are a kind of necessary fiction, simplifying the world in order to make it comprehensible, distorting in pursuit of understanding. This is, I think, a conclusion many philosophers come close to drawing (I have specifically in mind someone like Susan Oyama (2008)) yet retreat from doing so explicitly. Anyone who is an anti-realist about concepts, such as Gärdenfors, will at least implicitly adopt this view. It is one that F.A. Hayek would certainly have been amenable to. To confuse the fiction with the reality – to fail to perceive our inability to step outside the fiction – is once again to invite paradox. Paradoxes arise wherever we press too hard against the boundaries of our conceptual abilities.

To explore the paradoxes is to explore the boundaries. If the negative thesis of the chapter is that concepts are a kind of necessary fiction and that conceptual understanding is, *contra* Roger Penrose (1994), necessarily bounded, then the positive thesis is this: acknowledging and understanding our boundaries *extends* our conceptual reach. It absolves us of duties we cannot fulfill and allows us to see the value in (some) competing and seemingly mutually exclusive perspectives – mutually exclusive only because we cannot step outside our conceptual perspectives to resolve them.

If concepts are, by their nature, necessary fictions, then any theory of concepts, as itself a conceptual entity, can be no more. If concepts simplify and, by their simplification, distort the reality they simplify away from, then so will a theory of concepts. If concepts depend upon some essential inconsistency between what they represent and what they purport to represent, then a theory of concepts will be similarly dependent. Extreme care must be taken here: inconsistency in science or philosophy is generally considered a bad thing. An account that relies upon it must be approached cautiously, by small steps, if the resulting inconsistency is to be shown to be (to borrow a phrase from Chalmers (1996)) an innocent one. So I will attempt to do in this chapter.

First, I will re-frame the negative thesis with inspiration from a classic paper by Chalmers on consciousness. Thus framed, I will take it as a puzzle to break apart and re-assemble piece by piece, driving toward the conclusion that the inconsistency is both unavoidable and non-fatal.

The prize by chapter's end will be a powerful conceptual tool for toggling between competing pairs of perspectives on concepts, showing them first as representations, then as abilities; first as world-directed, then as self-directed; on the one hand private and personal, on the other public and shared; and so on. This is the natural conclusion of that idea first proposed in Section 2.7, of concepts being simultaneously representational and non-representational. Each half of the competing pairs of perspectives, taken on its own, will be indispensable; while any attempt at resolving the tension will raise difficulties, and full resolution will be impossible.

5.1 The Hard Problem of Concepts

> A scientific world-view which does not profoundly come to terms with the problem of conscious minds can have no serious pretensions of completeness. Consciousness is part of our universe, so any physical theory which makes no proper place for it falls fundamentally short of providing a genuine description of the world (Penrose, 1994, p. 8).
>
> The really hard problem of consciousness is the problem of experience.... Why should physical processing give rise to a rich inner life at all? It seems objectively unreasonable that it should, and yet it does (Chalmers, 1996, p. 6).

Proposal*:* as Chalmers has identified the *hard problem of consciousness*, so, too, one can identify a corresponding *hard problem of concepts.*

First of all I must argue that I am not simply creating the problem that I am then trying to solve. After all, if treating concepts as both representations and abilities is problematic, then, perhaps, one should insist that, ultimately, they are one or the other: i.e., that *one* account is primary, the other, at best, secondary, if not in fact dispensable. If, as Ruth Millikan argues, identifying concepts with categorization (as I have done in Section 3.4.4) yields much confusion, that is because it is fundamentally mistaken, and the critical aspect of concepts is their reliable ability to re-identify the same thing as the same thing – from many different perspectives, in many different contexts, by different modalities (see e.g. (Millikan, 1998, 2010))[1]. Concepts are not categories or representations, but rather *abilities* to form such things.

If allowing concepts both private and public aspects is problematic, then one might conclude that private (subjective) and public (objective) are two entirely different things, perhaps ones that mistakenly get muddled together; or one might conclude, as Fodor does, that concepts just are the sort of things that do not – cannot – vary in their contents from individual to individual. Alternatively, one could take the approach the psychologists favour and take the personal as primary. Likewise, if second-order concepts raise issues of paradox, why not establish a theory of concepts that deals *only* with first-order concepts, and then derive second- and higher-order concepts?

The reason why none of these strategies will work is, at heart, the same. Expressed one way, it is the problem of the ineliminable observer (Section 2.8); expressed another, it is Chalmers' target in his '95 paper: the problem of experience[2] and its enduring habit of mixing the seemingly objective (e.g., when we both measure the length of a certain stick, we get the same result, and we assume that, mistakes aside, others will get the same result, too) with the unavoidably subjective (the "rich inner life" in which that measurement is inevitably embedded).

Although we may, and should, conclude that there are first-order concepts, as soon as we try to say anything about them or even just observe them, the higher order intervenes! Although we may, and should, conclude that logically there must be a difference between our concept of e.g. dog and our concept of that concept (or between our concept of Fella and our concept of *that* concept) – the one clearly about a thing "in the world", the other not – the closer we attempt to examine

[1] This particular formulation is paraphrased from a conversation I had with Millikan at the SweCog Summer School in Mullsjö, Sweden, in August 2009. My response was and is that what makes the same thing the same thing is far from trivial; it's a major part of what needs explaining.

[2] ... By which I, and Chalmers, mean primarily conscious awareness: see e.g. (Chalmers, 1996, p. 20).

the matter, the more the lines become blurred; what at first we take to be simply in the world is inevitably coloured by our conceptual understandings and expectations of it (see Section 5.2.3). Much more will be said about conceptual expectations in Chapter Seven. Meanwhile, starting with first-order concepts, then deriving second- and higher-order ones, is simply not an option.

For much the same reason, it is not an option to start with concepts as some sort of non-representational entity – e.g., an ability to extract patterns, make associations, or form representations – then (optionally) derive concepts as iconic or symbolic representations, as many self-avowed anti-representationalists might attempt to do (e.g., (Brooks, 1991a,b; Perry, 1986; Gallagher, 2008)); for that would require stepping outside our representational perspective on concepts to get a perspective on concepts "as they really are". Of course, there is a sense in which the non-representational entity must logically be primary, to avoid grounding representations in representations and inviting a vicious regress; but there is an equally valid sense in which representations must be primary: because, whenever and wherever we look, they are there.

This is one way of understanding Gärdenfors' (2004) account of concepts as an intermediate level of explanation between association-based accounts and symbolic accounts of cognition: when one level of explanation is emphasized, concepts look more like non-representational abilities (or associations); when another, more abstract level, is emphasized, concepts look more like representations. But concepts, for Gärdenfors, like the theories about them, sit resolutely in the middle, beholden neither to one level nor the other.

Attempts to begin with concepts as public, objective entities likewise fall on stony or thorny ground, for where does private stop and public begin? My exercise of concepts in the public sphere is fundamentally shaped by my personal experiences of them, but my personal experiences are in turn shaped by my membership in society and, in particular, by my participation in a linguistic community. Furthermore, while the public may *seem* more objective, and is, indeed, less beholden to any particular individual perspective, there is no *a priori* reason to presume that it extends beyond the inter-subjective: at best, a useful approximation to "true" objectivity; at worst, something potentially just as biased as any particular individual perspective, if not more so. Consider the widespread expression of anti-semitic or anti-islamist attitudes, "everyone knows...", etc.

The bottom line is this: Chalmers claims that, "... a theory of consciousness should take experience as fundamental" (1996, p. 17). I think that a theory of concepts must do so, too, with all that entails in terms of e.g. subjectivity and perspectival bias.

For the conceptually reflective agent (see Section 4.1), concepts simply are part of experience. Concepts relate to consciousness through the medium of experience, which is why I think that the hard problem of consciousness and the hard problem of concepts are at least closely related if not one and the same. Experience gives rise to concepts, which in turn structure experience: the theme of Chapter Seven. For now, I suggest that the reader takes this passage from Chalmers and mentally substitutes the word "concepts" wherever he talks of "consciousness".

> The ambiguity of the term "consciousness" is often exploited by both philosophers and scientists writing on the subject. It is common to see a paper on consciousness begin with an invocation of the mystery of consciousness, noting the strange intangibility and ineffability of subjectivity, and worrying that so far we have no theory of the

> phenomenon. Here, the topic is clearly the hard problem – the problem of experience. In the second half of the paper, the tone becomes more optimistic, and the author's own theory of consciousness is outlined. Upon examination, this theory turns out to be a theory of one of the more straightforward phenomena – of reportability, of introspective access, or whatever. At the close, the author declares that consciousness has turned out to be tractable after all, but the reader is left feeling like the victim of a bait-and-switch. The hard problem remains untouched (1996, p. 7).

5.2 The Pieces of the Puzzle

One might accept, at least for sake of argument, that there *is* a hard problem of concepts. One might further entertain that that problem is the problem of (subjective) experience. But what does the "problem" really amount to? I suggest four things:

- **self-reference:** Theorizing about concepts is not only necessarily a self-reflective activity by an experiencing agent, it is self-referential in a way that raises certain logical difficulties. My conclusion is that this self-reference is, despite appearances, a *distorting* self-reference.
- **simplification:** The way concepts structure experience is to simplify it such that any original content is lost.
- **necessary fictions:** The illusion provided by concepts is that, in general, the original content is *not* lost. Concept pulls apart from referent only when we reflect on the matter, but the reflection is not the non-reflective use. Concepts possessed and employed non-reflectively make no such distinction.
- **paradox:** Experience places limits on our conceptual understanding by our inability to set that experience aside. Attempting to do so anyway, or failing to acknowledge the three points above, leads one into self-referential paradoxes. At the same time, we seem compelled, as a matter of pragmatic necessity, to treat experience, most of the time, as transparent (i.e., non-distorting): a necessary fiction or (mostly) innocent inconsistency.

The bottom line is that, as others have argued before me (notably Hofstadter (1979) in his analysis of the implications of Gödel's Incompleteness Theorem), completeness and consistency make uncomfortable bedfellows at best, and that at some point one must either choose between them or press no further.

5.2.1 Self-Reference

> There seems to be one common culprit in these paradoxes, namely self-reference... (Hofstadter, 2000, p. 21).

I wrote earlier about self-reference in sections 4.1.1 and 4.1.3. Self-reference is the reason why, according to Hofstadter, completeness and consistency are always in tension in any sufficiently expressively powerful system. Concepts need not, in any obvious way, be self-referential; but the activity of theorizing about them implicitly *is*. That activity critically involves self-reference on three different levels: the general nature of the enterprise, what the activity presupposes, and what it entails.

5.2.1.1 The General Nature of the Enterprise

> When the mind's focus is the focusing mind, new problems arise. The object and the instrument of the inquiry become one and logic is compromised. The mind is unable to decode itself or find its identity (Torey, 2009, p. 15).
>
> MIND, *n.*: A mysterious form of matter secreted by the brain. Its chief activity consists in the endeavor to ascertain its own nature, the futility of the attempt being due to the fact that it has nothing but itself to know itself with (Bierce, 1997).

As with consciousness studies, and indeed any sciences of the mind, theories of concepts take empirical study of the world – the more common domain of science – and turn it around, to focus attention on ourselves; and not just any aspect of ourselves, but that aspect that seems most essential to making us who we are: our minds. We can lose any other aspect of ourselves – an arm, or both legs, or even portions of our brain as happens in a stroke – and still feel, with apparent justification, that we are the same person. But if we lose our mind, then we really have lost ourselves.

Our minds are so privileged, and yet they resist a direct physical description. Rupert, given his concern for the "asymmetric relations between the persisting organismic portion of the purportedly extended cognitive system and the system's external portions" (2009a, p. 106) – the heart of his objection to Clark and Chalmers' extended mind hypothesis – would do well to remember this asymmetry, between our cognitive and biological identities. (I will argue for a version of the extended mind hypothesis in Section 5.2.3.)

The sciences of the mind are unique in this way, for the lines between observer and observed become significantly blurred: I cannot speak of mind without, implicitly if not explicitly, including *my* mind (for my mind is the mind I am presumably best acquainted with, and in any case am using to advance my theory). This line of thought – with its consequent limitations on what a science of mind can achieve – was explored by F.A. Hayek more than half a century ago, as noted in the opening quote to this chapter.

Although quantum physics theory has told us for some time that the observer cannot be removed from the equation, nevertheless much of science has proceeded on the basis that, for all practical purposes, the observer and any bias she introduces can be safely disregarded, and a "pure" objectivity achieved. (Hayek writes, "In order to be able to give a satisfactory account of the regularities existing in the physical world the physical sciences have been forced to define the objects of which this world exists increasingly in terms of the observed relations between these objects, and at the same time more and more to disregard the way in which these objects appear to us" (1999, p. 2).) At the least, observer and observed stand (or appear to stand) clearly apart. Such a distancing is not possible in studying concepts.

5.2.1.2 What Theorizing About Concepts Presupposes

> Theories of concepts focus neither on specific concepts nor on classes of concepts. They usually aim at characterizing the properties that are true of most, if not all, concepts – the general properties of concepts (Machery, 2009, p. 18).

Most, if not all, theories of concepts are as Edouard Machery describes: i.e., trying to set out the general properties of all concepts, to say what it is in general for something to be a concept (and, in the process, to establish a particular concept of concept). They are neither trying to restrict the class of concepts under consideration nor to hold "concept" to some restricted meaning, such as a technical term not intended to describe human thought generally – strategies that *might* attempt to avoid the self-reference noted in the previous section.

A further concern might be that whatever approach is taken, the theory is being put forward from inside a pre-existing conceptual structure, which it then purports to uphold. The broader the theory is, the greater the risk. This looks suspiciously – as it should – like the part (the theory of concepts, in this case) attempting to swallow the whole (all of concepts).

Consider Bertrand Russell's statement "the king of France is bald" (1905, p. 483) – which, intuitively, is hard to class as true or false because it implicitly presupposes that there *is* a king of France. (Russell read it to mean "there exists x such that x is the king of France and x is bald", on which reading the statement is clearly false. I wish to suggest instead that, although it presents itself as a statement with a truth value, its truth or falsehood is, because of its presupposition, impossible to resolve[34].)

In similar fashion, a theory of concepts presupposes, among other things, that the term "concept" is non-empty – Machery agrees that it is (2009, p. 15) – and that concepts form a kind whose properties are largely if not entirely in common. This Machery rejects (2009, p. 5); but what he describes as "kinds that have little in common" I think are better described as competing perspectives (and Machery offers no reason they should not be; it is simply not a possibility he considers). Again, we have plenty of experience of competing and apparently equally valid perspectives on what we agree must logically be the same thing: consider the description from physics of matter (or energy) as both wave and particle.

But the problem posed by the pre-existing conceptual structure may be much worse: in attempting to set out what a concept is in general (the "type"), the theory may seem to presuppose certain concepts (the "tokens") in particular, not least of them the concept of a concept itself (a token, which is also a type) – and not just their brute existence, but something of their nature. Most of the properties I listed in sections 3.1 and 3.2 – though probably *not* the property of evolvability – appear to fall into this category. (What would it mean, for example, for conceptual thought *not* to be systematic and productive, and still be conceptual?) As a general rule, the specification of a thing x should not include x in the specification, or the specification becomes circular (see Section 5.2.1.3).

[3] Ironically, on the same page, Russell offers "the king of England is bald" – which, at the time he wrote it, made a perfectly valid presupposition (that there was a king of England) and was, in fact, true. (At least, the king was going bald.) Today, whatever semantic status it is assigned will be the same as its more famous cousin.

[4] I also wish to suggest, *pace* Russell, that "the king of France" *did* and *does* denote something, just not the sort of something that it appears to denote: namely, a particular living individual. See the discussion of unicorns in Section 3.2.1.

One could argue that the pre-existing structure is mere scaffolding: incidental to the specification, not constituting a set of presuppositions at all. The presence of a general conceptual structure, or of particular concepts, in the specification is intended merely to evoke the necessary understanding, which is non-conceptual. But in that case, one might think that the conceptual content of the specification should, in principle, be dispensable once the specification is in place (as with true scaffolding); and it seems as if, in practice, it is not.

Here is what separates presuppositions from starting assumptions. After all, no theory of anything starts from a blank slate: one starts with some idea of the thing one is examining. (This makes the science fiction concept of encountering the "complete unknown" incoherent at best. If something were completely unknown, we would not even recognize it as something in need of an explanation.) But presuppositions lock one in in a way that other starting assumptions do not, precisely because those other assumptions *are* dispensable.

Consider the following thought experiment. Scientists discover an alien artefact. They speculate on how it is powered, what its intended function is, how old it is, and so on. Over time, all their initial hypotheses might be thrown out, even as to whether it is an artefact at all, as opposed to e.g. something that is or has been alive and has naturally evolved. Depending on what they discover, the very foundational principles of their sciences might get rewritten. What I am arguing could *not* change is some substantial portion of the underlying conceptual structure in which the scientists (as the rest of us) are embedded. Yet that underlying conceptual structure is precisely the intended target of a theory of concepts.

5.2.1.3 What Theorizing About Concepts Entails

> It is common practice to reject a justification out of hand if the argument that yields the justification is circular... (Brown, 1994, p. 406).

The heart of the problem with self-reference is this: the threat of what, in philosophical jargon, is known as *vicious circularity.* In ordinary reference, one has a thing referring, and a thing being referred to, and the two are not the same. In self-reference they are, or are substantially, or appear or claim to be, the same. If they are *precisely* the same, however, then one appears to have a problem with reference, because then not only is the thing referring to itself, but it is referring to itself referring to itself, and so on *ad infinitum*: a vicious circularity, or an endlessly receding target such as discussed earlier.

Consider a painting of a waterfall (a standard example of a representation), which refers to a (particular or generic) waterfall. The painting is not the waterfall, and no one is likely to mistake it for such. But consider a painting of a painting – and not just any painting, but that painting itself! If it captures a significant portion of itself, then we can recognize the self-reference: a painting of a waterfall that is also a painting of itself as a painting of a waterfall. But it cannot, on pain of infinite regress, capture itself perfectly. Indeed, long before that point it would cease to be recognizable as a painting of a painting of... a painting of a waterfall at all.

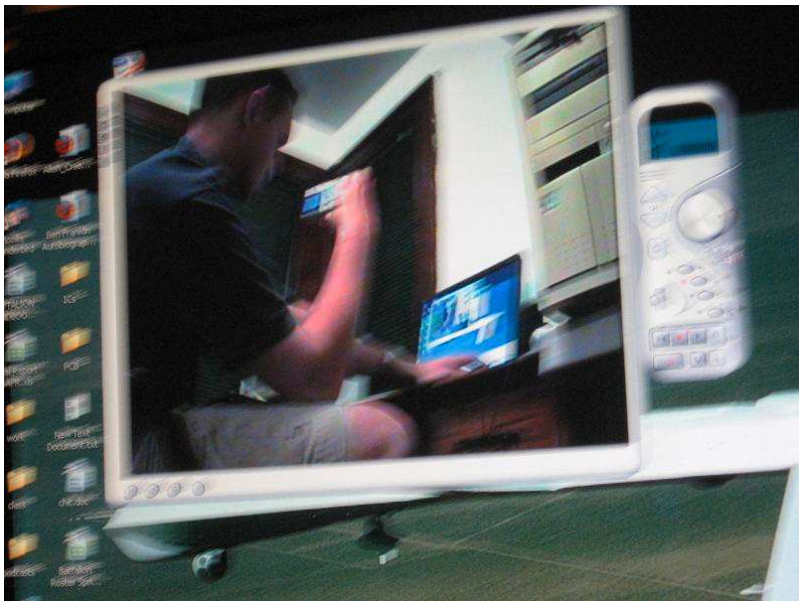

Figure 5.2: Photo by Luke Allen (http://lukeallen.org). Used by permission.

Consider likewise this photo of a photo (see Figure 5.2). The computer monitor can easily be distinguished on four levels: at the top-most level, as a picture on the desktop, as one item in that picture, and finally (just recognizable now!) one item in *that* picture.

Figure 5.3: Photo by Diggory Laycock (http://monkeyfood.com/blog). Used by permission.

The more the camera image captures itself, the more it becomes like Figure 5.3, an effect known as video feedback. Here, ten levels can be distinguished. As this process is pushed further and further, the resulting image becomes increasingly and rapidly abstract.

These images bear closely on the distinction Brown is at pains to make between acceptable and unacceptable forms of circularity, writing that "philosophers should be wary of rejecting justifications as viciously circular without examining the details of the argument that yields the justification even when there is a clear sense in which the argument is circular" (1994, p. 408). In particular, "the essential use of an hypothesis in the interpretation of a set of observations does not automatically prevent an empirical outcome that challenges that hypothesis" (1994, p. 409), and it is precisely this successful challenging of the hypothesis that reveals the circularity to be a non-vicious one (and the starting assumptions to be mere scaffolding). Both of these images are like the acceptable circularities – although, as the difference between them might imply, acceptable circularities exist along a continuum from the more acceptable (the counterpart of the first image) to the less (the counterpart of the second) to the viciously circular, with shades of grey in between. When "the essential use of a hypothesis" becomes sufficiently difficult to challenge – and one cannot challenge what one cannot make explicit – the circularity becomes to all intents and purposes vicious.

I want modestly to propose that theories of concepts exist along the same continuum. Push them too far the one direction, so that they attempt to capture too much of the conceptual structure from within which they are predicated, and they become like the snake swallowing its own tail (see Section 7.1.2). At best they are hopelessly confusing, their circular reasoning impossible to disentangle: video feedback reduced to noise[5]. At worst they are self-defeating, permitting the derivation of contradictory conclusions, such as the conclusion that the concept of not being potentially self-referential both is, and is not, itself potentially (or actually) self-referential (see Section 4.1.1).

Meanwhile, push theories of concepts too far the other direction, so that they fail to acknowledge enough of the pre-existing conceptual structure, and one risks valid accusations of Chalmers' bait-and-switch. At best they fail to explain what they claim to set out to; at worst they contradict the presuppositions implicit in the pre-existing conceptual structure. Those who would claim that concepts "just are abilities" or "just are (mental) representations" fall prey, I believe, to both.

5.2.2 Simplification

> The prime problem is that the information received by the receptors is too rich and too unstructured. What is needed is some way of transforming and organizing the input into a mode that can be handled on the conceptual or symbolic level. This basically involves finding a more *economic* form of representation: going from the subconceptual to the conceptual level usually involves a *reduction of the number of dimensions* that are represented... (Gärdenfors, 2004, p. 221).

As the previous section suggests, a good (as opposed to over- or under-reaching) theory of concepts simplifies in the pursuit of understanding; a theory of concepts that tries too hard to be faithful to what it is describing – i.e., to be complete in its account – by that very effort loses hold of its target; that is to say, it becomes unintelligible or inconsistent. So I have argued. If this is so, it should not be surprising, for it is, I believe, in the nature of concepts themselves.

[5] There is good reason why people find recursive structures notoriously hard to follow. Consider the progression of recursively embedded structures in the sentences: "The man left." "The man that the woman approached left." "The man that the woman that the child called 'Mother' approached left." "The man that the woman that the child that the dog bit called 'Mother' approached left." Linguists Morton Christiansen and Nick Chater write: "It is important to note that people's ability to process recursive constructions is quite limited. People produce only a very limited number of complex recursive constructions in naturally occurring speech, and this is reflected in the empirically documented difficulties that people experience when processing such structures" (2001, p. 6).

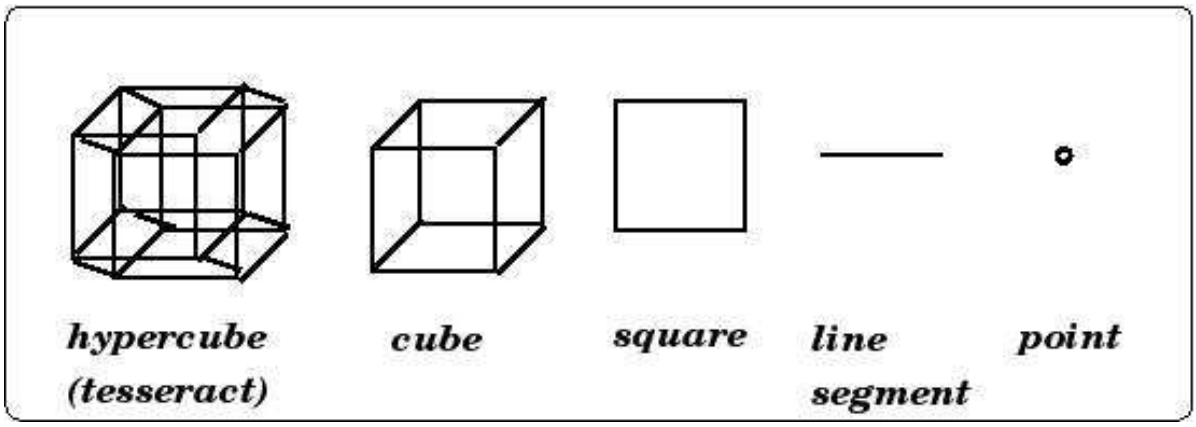

Figure 5.4: Compressing dimensions: an illustration from geometry.

I will expand greatly on this point in Chapter Seven, but for now suffice to say that what allows our conceptual abilities to do the kind of lossy compression I talk about in Section 4.3 is precisely their capacity to simplify the "too rich and too unstructured" "information received by the receptors". The price of this simplification is generally, if not universally, a "reduction of the number of dimensions that are represented": that is, the reduction in detail is not just quantitative but qualitative. The original version of something, in all its detail, cannot be recovered from the simplified version – however the simplification is done – without additional (external) input[6].

By way of analogy, consider Figure 5.4 and the way it doubly represents this reduction of dimensions. On the one hand a four-dimensional hypercube (or *tesseract*) has its three-dimensional counterpart in a cube, which has its two-dimensional counterpart in a square. Reduced to one dimension it becomes a line segment, reduced to zero dimensions a point. Note that the faces of a hypercube are cubes, the faces of a cube are squares, the sides of a square are line segments, and a line segment is composed of points. On the other hand, all of these are being represented on the two-dimensional surface of a piece of paper, which renders the hypercube borderline incomprehensible and seriously distorts the cube. Only the square appears undistorted. (The line segment must be given width to be visible at all, and the point must be given both a minimal length and width.)

5.2.3 Necessary Fictions

> Philosophy is done on the basis of a noble lie, a necessary fiction, namely the belief that a thesis can be expressed unambiguously and evaluated conclusively. All the evidence is against it, but one must believe it (Cupitt, 1980, p. 32).

A "necessary fiction" I will define as something that we recognize, logically, cannot be (quite) true, but that we cannot do without. When ultimate truth outstrips our capacity for understanding, *and we can recognize that*, then necessary fictions must suffice.

A fiction, to be a *good* fiction, needs to be plausible; normally at least, it should have a "could happen" or "could have happened" feel to it. Fictions can be more or less truthful; perhaps, too,

[6]This incompatibility between exact precision and comprehendibility is captured by Zadeh's *incompatibility principle*, presented in (Novak, 2005, p. 343) as: "... The increase in exactness leads to an increase in the amount of information whose relevancy then decreases until a point is reached after which the preciseness and relevancy are mutually excluding characteristics".

truths can be more or less fictional, or at least more or less true. Most of the time our approach to truth is pragmatic rather than dogmatic. Many truths may never be questioned; some may not be open to question; but this is different from placing any truths beyond question.

Consider the following exchange between the king of one-dimensional Lineland and Square, a resident of two-dimensional Flatland, from Edwin Abbott's 19th Century novel. Part of Abbott's intended moral is that we, too, for all our apparent superiority over the Lineland king or Square, should take our understanding and imagination to be bounded. (We have an advantage over the king or Square, for we can logically conceive of things like hyperspheres and hypercubes[7], even while we cannot visualize them and cannot construct them – only their three-dimensional projections or *shadows*. This logical conceivability is hugely important, for, as I suggested already in Section 2.7, it is what allows us to understand concepts as non-representational abilities even though, when we observe them, they are always representations. It is what allows us to draw a distinction between "seemingly are" and "logically could or must be".)

> "Not so," replied I; "besides your motion of Northward and Southward, there is another motion which I call from right to left."
>
> **KING**. Exhibit to me, if you please, this motion from left to right.
>
> **I**. Nay, that I cannot do, unless you could step out of your Line altogether.
>
> **KING**. Out of my Line? Do you mean out of the world? Out of Space?
>
> **I**. Well, yes. Out of YOUR World. Out of YOUR Space. For your Space is not the true Space. True Space is a Plane; but your Space is only a Line (Abbott, 1885, p. 97).

So the first necessary fiction, for the king and for Square – as, Abbott suggests, it is also for us – is that the world-as-perceived just *is* the world. That is to say, the perception is, at least for the most part, transparent: not simplifying, not distorting, neither adding anything nor taking anything away. What goes for perception goes likewise for the concepts that structure that perception.

The second necessary fiction is that concepts and referents reliably and unproblematically pull apart.

5.2.3.1 Separating Mind from World[8]

To return to my earlier example, no one would confuse the painting of a waterfall with the waterfall itself. On first blush, one would never confuse one's concept of a dog (in general or of one dog in particular) with the dog itself, either, nor conflate the two together; and yet, as I hope to convince you, this is what conceptual agents do all the time when they are employing concepts non-reflectively, when they are not thinking about them and are just getting on with using them. Making such a distinction is a conceptually higher-order activity, one that may seem to require beliefs about beliefs. It is only when we reflect upon things that concept and referent pull apart; and even there, if we examine matters too closely, we run into difficulties finding the line between the two.

[7] For an excellent fictional introduction to hypercubes, see (Heinlein, 1983).
[8] Much of this and the next subsection appear in (Parthemore, 2011).

Recall the class/instance distinction that I made in Section 2.1, and the suggestion there that instances could also be understood as classes, because of the way they generalize (and at the same time, as the previous section should suggest, simplify) over specific experiences. (I will return to this distinction again in Section 6.2.) It is time to press that point further and, in doing so, undermine the strict distinction between concepts and non-concepts[9].

Let me use again the example of your pet dog Fella. Upon any specific Fella encounter, you bring a great deal of conceptual expectations to bear, whether you are reflectively aware of doing so or not (and, in the usual circumstances, you probably will not be). What you experience is an object, with all the expectations of e.g. object permanence that you have had from a very early age[10]; but not just any object: a dog, with all your expectations about dogs (that *ceteris paribus* they *do* bark, that they *never* purr); but not just any dog: your dog Fella, who crawls into your bed every night, whom you've raised since he was a puppy, whom you took to the vet's last Tuesday for deworming. Whatever it is, the referent of "my dog Fella" is not, or is never just, the thing-in-itself, stripped of all conceptual shading. To recognize Fella *as* Fella on any particular occasion – or, more minimally, as a dog, or more minimally yet, as an object – is already to have passed beyond the possibility of any strictly in-the-moment, strictly non-conceptual experience. So I will argue in Chapter Seven, where I will explore further the nature of non-conceptual mental content. For now, suffice to say that logic implies the necessity that the referents of *all* concepts are (or may be) conceptually coloured[11], further blurring the line we already started to blur at the start of Chapter Four.

And yet our understanding of the distinction between x and our understanding of x is so basic, for most of us, as perhaps to appear trivially obvious. It is surely a prerequisite to successfully navigating the so-called false belief task (Wimmer and Perner, 1983). It implies, at the least, implicit beliefs about beliefs, if not in fact the sort of explicit meta-beliefs that Davidson sees as being necessary for surprise (1987). It is flawed: not necessarily because of any problem with the distinction itself but because of the assumption – seemingly unavoidable – that we can, at least most of the time, make it reliably, without interference from conceptual prejudices. It creates an inconsistency, between a seemingly conceptually uncoloured "external" world and one we may logically conclude to be potentially touched, *wherever we look*, by our conceptual influence.

5.2.3.2 The Extended Mind

> Where does the mind stop and the rest of the world begin (Clark and Chalmers, 1998, p. 7)?
>
> Natural cognitive systems are enormously subtle and complex entities in constant interaction with their environments. It is the central conjecture of the Dynamical Hypothesis that these systems constitute single, unified dynamical systems (van Gelder and Port, 1996, p. 11).

[9] Recall that I have already, in Section 4.1.1, shown the distinction between first- and higher-order concepts to be problematic. This is really the same problem in a slightly different guise.

[10] Jean Piaget, who coined the term, famously located this ability at nine months (1954); more recent research (e.g. (Ballargeon, 1987)) has shown reliable evidence for an expectation of object permanence at less than half that age.

[11] Empirical evidence does as well. Hemeren (2008, p. 150) cites evidence from (Grill-Spector and Kanwisher, 2005) of high-level categorization influences on early visual processing: strong evidence that vision (and by implication all our sense modalities) is *not* conceptually transparent!

If all of this sounds vaguely reminiscent of the debate over the *extended mind hypothesis* (Clark and Chalmers, 1998; Clark, 2008) (EMH), it should. If anything bears the "mark of the cognitive" – as noted earlier, one of Adams and Aizawa's stock expressions (2001; 2008) – surely concepts do. If conceptual influence bleeds "all the way out" into the world, then cognition plausibly extends (in some meaningful sense) into the world as well. One need not maintain – as McDowell is often read, and as the Dynamical Hypothesis could be read (see Section 7.1.1) – that the world we encounter is *fully* conceptual, only that there is no part of that world that is fully or reliably free of the conceptual touch. Such cognitive tentacles into the world are all that the EMH requires.

I happen to share Adams and Aizawa's, and Rupert's (2004; 2009a; 2009b) concerns about "cognitive bloat". The moral of the story may be that a little extended mind goes a long way; and all this requires, I believe, is a sufficiently flexible boundary between mind and world, one that shifts over time in a way that Clark talks about in ***Supersizing the Mind*** (2008) (what I take to be the best strategy for advancing a version of the EMH). All of the standard objections to the EMH depend on a clear and relatively fixed boundary between mind and world – and that, I have argued already in Section 1.5, one cannot take for granted without certain metaphysical assumptions.

5.2.3.3 Innocent Inconsistencies

There are two ways one could misinterpret the Donald Cupitt quote that began this section. (Cupitt is a philosopher and theologian, as well as Anglican priest.) One is that he is "dissing" philosophy, in a way that Wittgenstein can sometimes be read. The other is that he is intending *only* to comment on philosophy. His target is indeed **"**the belief that a thesis" – *any* thesis – "can be expressed unambiguously and evaluated conclusively". I hope to have convinced you by this point that such a goal, though "noble", is, by virtue of the limitations on our conceptual abilities, impossible.

An "innocent inconsistency", then, is one we can live with. We recognize, intellectually, that it's there; but, for the most part, it does not interfere with our scientific or philosophical theorizing or our ordinary day-to-day lives. Unlike inconsistencies in logic, where one inconsistency renders the whole system inconsistent – where one falsehood can be used to derive anything as true – it does not bring down the entire system.

5.2.4 Paradox

> Epimenides was a Cretan who made one immortal statement: "All Cretans are liars." A sharper version of this statement is simply "I am lying"; or, "This statement is false" (Hofstadter, 2000, p. 17).
>
> A paradox may be defined as an apparently valid argument with apparently true premises and an apparently false conclusion (Williamson, 1996, p. 22).

To recap: concepts are necessary fictions, enabling us to understand the world at the same time that they distance us from it. The paradoxes one reveals when one pushes those fictions too far serve to reveal the very real limitations on human conceptual cognition. That is what I think these paradoxes ultimately are: human cognition pushing up against its boundaries; and that is why, I suspect, we tend to find paradoxes intuitively meaningful. Understanding the limitations of understanding – knowing as much as we can, not just about what we do not know but about what (perhaps) we cannot know – is, I believe, one of the most fruitful forms of understanding. At

the least, cognitive science and AI need to be aware of these limitations; at best, they can learn to exploit them.

Although fine so far as it goes, Timothy Williamson's definition of paradox will, I fear, admit things I prefer not to call paradoxical and potentially reject things I would. I prefer the following: a paradox is any set of circumstances[12] such that one has equally valid and compelling grounds for concluding both a proposition p and its negation $\tilde{}p$, such that *it is impossible on pain of contradiction to choose between them.* (It cannot be enough merely that one's judgment could go one way or the other, even if such circumstances are occasionally described, in a loose way, as paradoxes.) This, I believe, captures many if not most the familiar paradoxes of logicians and philosophers and best gets to the heart of the matter.

Russell's Paradox is, as Russell himself noted (1908), a variation on the Epimenides or Liar's Paradox, as is Grelling's Paradox (see Section 4.1.1), as are uncountably many other variations. What all have in common is, as Russell and Hofstadter among others have noted, a vicious self-reference: in particular, the seeming ability of a part of some whole to capture the whole. In the case of the Epimenides Paradox, it is the way "all Cretans" includes the speaker himself and so implicates all the speaker's statements, including the statement "all Cretans are liars" itself. This is what leads Russell to offer the rule: "Whatever involves *all* of a collection must not be one of the collection" (1908, p. 225).

Let me try to be clearer about what exactly is going on here. The sort of self-reference behind these paradoxes admits of two possibilities, precisely the two raised earlier: *either* one has an eternally receding target *or* one eternally flips between two exclusive and opposing perspectives. For example, with Russell's Paradox, the set of all sets that contain themselves contains itself, contains itself containing itself, contains itself containing itself containing itself, and so on. This is like the video camera *perfectly* capturing its own image. Meanwhile, for the set of all sets that do not contain themselves as members, one can speculate that it does not, in fact, contain itself; but by that very choice, it *does* contain itself. But if it does contain itself, then by *that* choice, it does not. Compare what Hofstadter writes about the Epimenides Paradox: "It is a statement that rudely violates the usually assumed dichotomy of statements into true and false, because if you tentatively think it is true, then it immediately backfires on you and makes you think it is false. But once you've decided it is false, a similar backfiring returns you to the idea that it must be true" (2000, p. 17).

5.2.4.1 Escaping Paradox

Given a paradox such as the ones above (one that really is a paradox, and not a trivial case of faulty reasoning), a number of obvious strategies present themselves.

1. One can disallow the problem cases that allow the paradox to arise: that is, they amount to faulty reasoning after all. Here there are two possibilities:

   (a) One can allow the problem cases to arise but then reject them *post hoc* as simple contradictions. This I take (and Jolley (2007) takes) to be Russell's solution (1908).

[12] Paradoxes typically are expressed in narrative form as a proposition or set of propositions. They need not be, however: consider the visual paradoxes created by M.C. Escher. The really critical point is the "equally valid and compelling grounds for...", a formulation I owe to Ron Chrisley (personal communication).

(b) One can prevent them from arising in the first place. Jolley reads this (correctly, I think) as Wittgenstein's preferred alternative to Russell.

2. One can grasp both horns of the dilemma and permit the derivation of a contradiction. This is generally (and wisely) considered a bad thing, although Hofstadter (1979) offers some intriguing thoughts from Zen Buddhism why this need not necessarily be so.

3. One can allow the problem cases as genuine paradoxes and conclude either that:

   (a) Our present perspective is not broad enough to grasp how the apparent contradiction can be resolved (*cf.* the discussions over whether matter and energy are best understood either as wave or particle).

   (b) Or: our perspective *by its nature* may *never* be broad enough to grasp how the apparent contradiction can be resolved.

Contra Russell or Wittgenstein, I want to hold that there *is* something meaningful to paradoxes. At the same time, while permitting the derivation of a contradiction may be useful meditationally, scientifically it is not. Therefore I wish to retain the problem cases, and furthermore conclude that some of them, at least, will never resolve away.

I take my cue here from Hofstadter, who thirty years ago, with his talk of "tangled loops" and self-referencing systems (1979), attempted to bring paradox to the heart of cognition itself. There is, I have claimed, an inescapable problem both with reflecting on how we reflect upon our concepts, which threatens an eternally receding target (at what point do we grasp the concept itself?); and with reflecting on how we do *not* reflect upon our concepts, which threatens an eternal oscillation between two competing and contradictory points of view: that concepts both must, and cannot, be representational.

5.2.4.2 Slippery Slopes

> A typical vague property is *to be a small natural number.* Can we imagine all the small natural numbers? Clearly, 0 is small, 1 is small as well, etc. But where does this sequence finish? The only sure fact is that there exists a number, for example 1,000,000,000, which is *not small* (Novak, 2005, p. 343).

I should mention, at least in passing, another set of paradoxes that meet the definition I offered at the start of this section but that do not (at least in any immediately obvious way) involve self-reference: namely, the Sorites ("heap") Paradox, the Surprise Examination ("prediction") Paradox, and their variations[13]. All take an apparently valid premise and, by a series of apparently valid steps in reasoning, arrive at a conclusion that contradicts the premise[14]. Consider:

1. A billion grains of sand is a heap and a single grain of sand is not.

2. If a billion grains of sand is a heap, then 999,999,999 grains of sand is a heap.

3. If 999,999,999 grains of sand is a heap, then 999,999,998 grains of sand is a heap.

[13] It should be noted that the Surprise Examination Paradox is frequently (e.g., by (Fitch, 1964); see also (Chow, 1998)), understood as involving an implicit self-referential assumption.

[14] This relates them to a wider set of problems that are not, strictly speaking, paradoxical. I once assisted a friend on the programming portion of his master's thesis, an AI program exploring syllogistic reasoning. By a series of carefully chosen syllogisms we led the user of the program from an innocuous starting point – some point about abortion that one might readily agree to – to the conclusion that all men should be systematically exterminated.

4. ...

5. If two grains of sand is a heap, then one grain of sand is a heap.

6. Therefore one grain of sand is a heap.

Another, slightly more complicated, chain of syllogisms can be constructed for the Surprise Examination Paradox.

I do not intend any sophisticated treatment of these paradoxes; others have done that far better than I might. What is relevant for my purposes, though, is the way all of these paradoxes can be interpreted as resulting from inexact knowledge – the way knowledge generally involves a certain "margin of error" – and the distinction some would draw between what we know and what we know that we know; Williamson (1992) takes this approach. In the case of the Sorites Paradox, the inexact knowledge arises from a vague concept (**HEAP**) masquerading as a clearly defined one (Williamson, 1992, 1996). Putting this another way, one has a continuum (from heap to non-heap) masquerading as a binary distinction.

5.2.4.3 Binary Distinctions and Underlying Continua

> A CP [categorical perception] effect occurs when... a set of stimuli ranging along a physical continuum is given one label on one side of a category "boundary" and another label on the other side.... In other words, in CP there is a quantitative discontinuity in discrimination at the category boundaries of a physical continuum... (Harnad, 1990b, p. 3).

But the idea of continua underlying binary distinctions should sound quite familiar: it has been one of the recurring themes of this work[15]. So I have written e.g. of the continuity between symbolic and sub-symbolic or non-symbolic (Section 2.6.3), between non-conceptual and conceptual mental content (Section 3.4.4.2), between low-level (directly sensorimotor-based) and high-level (abstract) cognition (Section 3.4.4.3), between first- and higher-order concepts (Section 4.1.1), between non-human animal and human cognition (Section 4.4), between vicious and non-vicious circularities (Section 5.2.1.3), and finally between conceptual agent and environment (Section 5.2.3.2) – a point that shall come up again quite prominently when I introduce the enactive approach in Section 6.1.2.

All of this is as one would expect if, as suggested in Section 3.4.3.1, categorization – the means by which we separate what is x from what is not x – is intrinsically conceptual, meaning that, *contra* most natural kinds accounts, categorization does *not* map categories of concepts to categories in the world; and, *so far as concepts are concerned*, the world is continuous, while it is concepts that are discrete: an idea I freely borrow from the research area of *categorical perception*. Examine the boundaries between concepts too closely and the sharp lines dissolve back into continua, the better to reflect the world they abstract away from. Frege wanted all concepts to be non-vague. If I am right, then (mirroring a point made nearly a century ago by Russell for language (1923)), all (or nearly all) concepts are ultimately vague, and the inexactness of conceptual knowledge is parasitic on this.

[15] *Cf.* the substitution of multiple for binary values in *fuzzy logic*, which, also, is trying to get at continua underlying binary distinctions.

5.3 Why Cognition *Is* Bounded: Reflections on Penrose

The argument of this chapter, in brief: concepts need not be self-referential, but theories of concepts, as higher-order conceptual entities, always, at least implicitly, are. Push that self-reference too far and it becomes a vicious self-reference. Concepts, by their nature, abstract away from the particulars of any given context and, in the process, simplify, reducing the "too rich" structure of low-level cognition (which may already be [proto-]conceptually touched). At the same time that concepts appear to hold up a non-distorting mirror to the world, we can logically conclude that this is, at best, a necessary fiction, and that concepts are such fictions as well: not ultimately true to a pre-experiential world, but *true enough*. Press against the boundaries of conceptual understanding – or, worse yet, ignore them – press the self-reference, press the simplification, press the necessary fictions, and one comes up against paradox.

My position is, needless to say, tendentious, and many arguments have been raised against positions much like it. Most interesting and relevant for the present discussion is Roger Penrose's argument that human cognition is non-computational and not bound by Gödel's Incompleteness Theorem[16]. For my purposes I need not show that human cognition is algorithmically describable (what Penrose means by "computational"), only that there is strong reason to think that, regardless, it *is* logically bound by the same limits that Gödel's Theorem puts on formal computational systems. Before I explain where I think Penrose is wrong, though, I need to say something about where I think he is right.

5.3.1 Gödel's Revenge

> It rapidly became accepted as being a fundamental contribution to the foundations of mathematics – probably the most fundamental ever to be found – but I shall be arguing that in establishing his theorem, he also initiated a major step forward in the philosophy of mind. Among the things that Gödel indisputably established was that no *formal* system of sound mathematical rules of proof can ever suffice, even in principle, to establish all the true propositions of ordinary arithmetic (Penrose, 1994, p. 64).
>
> Mathematical statements – let us concentrate on number-theoretic ones – are about properties of whole numbers. Whole numbers are not statements, nor are their properties. A statement of number theory is not about a statement of number theory; it just is a statement of number theory. . . .
>
> Gödel had the insight that a statement of number theory could be about a statement of number theory (possibly even itself) if only numbers could somehow stand for statements. The idea of a code, in other words, is at the heart of his construction. ...This coding trick enables statements of number theory to be understood on two different levels: as statements of number theory, and also as statements about statements of number theory (Hofstadter, 2000, p. 18).

As noted earlier, Russell saw paradoxes as nothing more than contradictions, and so set about banishing them – not just from set theory (Russell, 1908) but from mathematics more generally. His *magnus opus*, ***Principia Mathematica*** (PM), written with Alfred North Whitehead, set out to derive a complete (and, of course, consistent) set of mathematical truths from a set of well-defined and universally accepted axioms and inference rules: the Holy Grail of mathematical

[16] There actually are two, closely related theorems, originally published in (Gödel, 1931).

foundationalism. The story is told very well by Hofstadter (2000, pp. 19-24), himself a trained mathematician.

Gödel is generally credited with having established two significant – some might say crippling – limitations in PM and *verwandte Systeme* ("related systems"). First, there are statements expressible within the framework of PM, and indeed of *all consistent axiomatic systems* of sufficient expressive power, that are expressible within those systems yet cannot be proven within those systems if the systems are not to be self-contradictory. Such a statement for any axiomatic system T – there may be several or many – is commonly known as a Gödel sentence of T. The Gödel sentence of PM might be translated as, "This statement of number theory does not have any proof within the system of *Principia Mathematica*" (Hofstadter, 2000, p. 18). The problem cannot be resolved by resorting to a larger, more comprehensive axiomatic system because, although the original Gödel sentence will not be a Gödel sentence within the new system, something else will be: i.e., the problem involves an *essential undecidability.*

Second, and relatedly, the consistency of PM or any other axiomatic system cannot be established within that system, for were it to prove its own consistency, it would by that act become inconsistent. (It is trivial to show that any *inconsistent* system can prove its own consistency, by virtue of being able to prove anything.) As Hofstadter writes, "the final irony of all is that the proof of Gödel's Incompleteness Theorem involved importing the Epimenides paradox right into the heart of *Principia Mathematica*, a bastion supposedly invulnerable to the attacks of strange loops!"(2000, p. 24)[17]

A few caveats are in order. A discussion of higher-order mathematics or logic is outside the scope of this work. The correct reading of Gödel's Theorem is not entirely uncontroversial even within mathematics: e.g., as to its impact on the so-called Hilbert's Program, David Hilbert's attempt to establish a secure foundation for mathematics. Not surprisingly, attempts to apply it outside mathematics (and there have been many[18]) have proven even more controversial.

Still, in such company as Douglas Hofstadter, I share Penrose's intuition that Gödel's Theorem is critically important to philosophy of mind. Most critically, it gets us thinking in a more rigorous way about the nature of those thought processes that make mathematics possible, as well as about the nature of knowledge, understanding, and proof. Too, it gives concrete expression to our strongly intuitive yet often fuzzy notions about paradox.

I agree with Penrose on many other points as well: e.g., the incapability of Turing machines *as idealized mathematical entities* to capture something essential about e.g. human understanding or cognition[19]. Indeed, I think Penrose would have no issue with most, if not all, of what I say

[17] As I noted earlier (Section 5.2.1.2), "strange loop" is Hofstadter's coinage for the kind of self-reference that gives rise to paradox.

[18] . . . Such as attempting to use it to disprove the existence of God, prompting the joking "response" that "God disproves Gödel". Interestingly, Gödel himself (1995) formalized St. Anselm's ontological argument for the existence of God.

[19] On the other hand, I find statements like the following just wrong: "Although a modern computer's detailed internal construction is very different from this (and its internal 'working space', though very large, is not infinite like the Turing machine's idealized tape), all modern general-purpose computers are, in effect, actually universal Turing machines" (Penrose, 1994, p. 66). Universal Turing machines precisely are abstract, idealized entities, operating independently of any influence of environment and any consideration of time. All real-world computers, on the other hand, are embedded in a particular environment and embodied (however relatively weakly) in a particular physical form, in ways that non-trivially matter. Contrary to Penrose's confidence that computers do not make mistakes, computers are not, ever, in practice so reliable – as anyone knows who has ever lost an afternoon's work because e.g. the processor overheated and shut down the computer, or the hard drive "died", or somebody spilled coffee in

above. Why, then, does Hofstadter conclude – as I would like to – that Gödel's Theorem reveals a fundamental limitation in human understanding and cognition, while Penrose comes to nearly the opposite conclusion?

5.3.2 Parting with Penrose

> If, as I believe, the Gödel argument is consequently forcing us into an acceptance of some form of viewpoint C[20], then we shall also have to come to terms with some of its other implications. We shall find ourselves driven toward a *Platonic* viewpoint of things. According to Plato, mathematical concepts and mathematical truths inhabit an actual world of their own that is timeless and without physical location. Plato's world is an ideal world of perfect forms, distinct from the physical world, but in terms of which the physical world must be understood. It also lies beyond our imperfect mental constructions; yet, our minds do have some direct access to this Platonic realm through an "awareness" of mathematical forms, and our ability to reason about them (Penrose, 1994, p. 50).
>
> The concepts that we "know" in this Platonic sense are things that are "obvious" to us – they are things that can be reduced to a perceived "common sense" – yet we may not be able to characterize these concepts completely in terms of computational rules (Penrose, 1994, p. 54).
>
> . . . The notion of "unambiguous definition" cannot itself be unambiguous. Similarly, the notion of "unassailable truth" cannot itself be unassailable (McCullough, 1995, p. 5).

To be clear: I part company with Penrose *not* over his conclusion that human understanding involves something non-computational or non-algorithmic – though I suspect he is in some substantial measure wrong there, and that his subsequent resorting to quantum gravity effects in the microtubules of neurons to explain consciousness should suggest a wrong turn somewhere**.** (I am in good company. As Rick Grush and Patricia Churchland note, Penrose should *not* be misread as limiting "algorithmic" to "involving explicit use of an algorithm", which would, indeed, rule out much of human cognition and understanding. He means to include *anything* that can, in principle, be captured by an algorithmic description. "'Non-algorithmic' in the sense Penrose quite reasonably intends is, therefore, a very strong constraint; so strong in fact, that whether there exist any physical systems in the real world whose behavior is non-computable in this strong sense remains very much an open question" (Grush and Churchland, 1995, p. 190).) My argument is rather with how he comes to that conclusion, because if he is right, and human cognition is *not* bound by the logical boundaries that Gödel's Theorem sets – or worse, may not be logically bounded at all[21] – then I risk losing one of the main conclusions of this chapter.

I do not contest his conclusion as more narrowly expressed, that "human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth" (1994, p. 76). Indeed, I think he is probably right on that score. I think he is right as well that the way we do mathematics is significantly revealing of the way our thoughts work more generally. What I reject is his conclusion that the methods mathematicians are using *must nonetheless be knowably*

the wrong place. Their operation is determined in interaction with their environment, as is ours.

[20] "*C*. Appropriate physical action of the brain evokes awareness, but this physical action cannot even be properly simulated computationally" (Penrose, 1994, p. 12).

[21] Of course an *individual's* understanding will be limited by the physical limitations of her mind and brain, as well as the practical limitations of her education, but it is logical boundaries that Penrose and I are interested in.

sound (and therefore non-algorithmic). As Grush and Churchland note (1995, p. 193), those methods could conceivably be unsound but benignly so: i.e., in a way that does not affect its basic conclusions. Likewise they could be, not *knowably* sound, but reasonably confidently believable (Grush and Churchland, 1995, p. 195), which strikes me as the likeliest answer. A lot turns, as noted earlier, on what we *know* that we know.

To be up-front about my own intuitions (see again Section 1.5): I am not a Platonist, and I would be inclined to question any conclusion that pushed me toward Platonism. There are, for me, no "shadows on the cave wall": no shadows, no cave, no wall. Like Kant, I believe we have no epistemic access *whatsoever* to a pre-experiential (self or) world; we can conclude, only, that logically such a world must be there, constantly constraining our experiences and actions. But if Penrose could accuse me of letting my metaphysical leanings colour my thinking, then I could likewise accuse him of letting his incline him to accept "obvious" truths at face value (and to rely on them, much of the time, for his argument)[22]. After all, not only do "obvious" philosophical intuitions vary widely between cultures – as Stephen Stich has shown to devastating effect (Weinberg et al., 2001; Stich, 2006) – even within a culture, intuitions can, and do, differ, even among mathematicians (Penrose acknowledges but downplays this) and hard-core empirical scientists. How else can one explain the contrast between the quote attributed to Einstein in Chapter One and this from Stephen Hawking: "... There are grounds for cautious optimism that we may now be near the end of the search for the ultimate laws of nature" (1988, p. 172)?

In the end and borrowing a page from Daryl McCullough, I conclude that, *pace* Penrose, what we know that we know may be somewhat less than what we think that we know or what we actually know. More to the point, what we know (regardless of whether we know that we know it) depends critically on what we do *not* know: i.e., it depends on our knowledge being incomplete. Once again, completeness and consistency sit poorly together.

The details of my argument against Penrose are not worth going into here. They can be found in Appendix A.

5.4 The Toggling Effect

> ... We may take an object and just by focusing on it we notice almost at once that *it* (the content component) begins to recede and become overlaid by the nonthematic sensation that the whole experience is our own doing. However, this same sense of self-contribution, too, begins at once to fade, allowing the attention to swing back once more to the object in focus, from there to fade in turn, accentuating the self-sensation once more before the attentional pendulum swings back to the object again (Torey, 2009, p. 112).

The negative conclusion of this discussion is that there is no reason to think that informal systems can handle joint completeness and consistency any better than formal ones can. Informal systems are just as effective at deriving paradoxes. We should safely conclude (even if we cannot know that we know) that there are genuine limits to conceptual understanding of anything, not least of all ourselves.

[22] This is the heart of my argument, in Section 1.5, against Adams and Aizawa's, and Rupert's, rejection of the extended mind hypothesis.

The positive outcome is the prize I promised at the beginning of the chapter: the "powerful conceptual tool for toggling between competing pairs of perspectives". The germ of the idea was first introduced in Section 2.7 and has been gradually developing since then. It is now time to restate and elaborate upon it.

To wit: attempting to come to terms with concepts and their nature gives rise to two contrasting views, both logically necessary and both logically limited. One threatens us with an eternally receding target (reflecting on our reflections to try to get at the actual fact of the matter), the other an eternal oscillation between two opposed and contradictory positions (reflecting on what is going on when we are not reflecting on what is going on). Furthermore, to the extent we examine the nature of our concepts, we cannot help but toggle back and forth between the two: we are, in effect, caught within another oscillation.

I take this to be essentially the same point Torey is making when he writes about an "oscillation of attention" (2009, p. 39). Where he and I would differ is that he thinks the human mind is capable of escaping this oscillation to see things "as they really are", and I think it is not; as well, he sees the "content" view as primary and the "self" view as secondary, whereas I think which one is primary is solely dependent on which is in the present foreground of attention. This follows from what I said in Section 2.8.

Putting both views within the same theory of concepts creates an (apparent) inconsistency, but it is an innocent one, born of limitation of perspective. Either view taken on its own is *more* problematic than the two views taken together. The lesson to be learned is that, in place of a complete and consistent account of concepts, our goal should be a working understanding that presses against the boundaries of our ability to know ourselves.

These two views are (at least mostly[23]) consistent unto themselves, so long as the boundary between them is not explored too carefully. Depending on the discussion in question and on the context (psychology versus philosophy of mind versus cognitive science versus AI) they can take different forms; but all the forms they can take relate back to these two, basic views. Call them the "observational" and "operational" perspectives. On the one hand, concepts are being employed by an agent non-reflectively, without consideration of the concepts as concepts (the *operational* context); on the other, concepts are being employed reflectively, where the object of awareness is the concepts themselves (the *observational* context). Logical necessities apply to the operational context, conceptual necessities to the observational context. With this in mind, consider that, as Gödel brought paradox to the heart of ***Principia Mathematia*** (see (Hofstadter, 1979, p. 24)), so, too, a paradox can be seen to lie waiting at the heart of any theory of concepts, expressed in the following two propositions:

> The concept of concept must be defined in terms of itself.
>
> The concept of concept cannot be defined in terms of itself.

The first proposition is *conceptually* necessary, the second *logically* necessary.

[23] There is nothing *prima facie* to prevent the possibility of further pairs of contrasts being revealed *within* each view.

A lot of disagreements among contemporary accounts of concepts line up remarkably well on one or the other side of this *operational/observational* divide, as Table 5.1 suggests. So for example, when we reflect on concepts *as* concepts, it may seem that they *just are* representations and symbolic language *just is* the natural language to use to describe them. And yet, when we attempt to consider logically how we must get on with employing concepts non-reflectively, it appears that they are not symbols or representations at all. Likewise, when we reflect on concepts, they appear as static, discrete, abstract mental objects, even though logical consideration leads us to think that they must be dynamic as opposed to static, continuous with other concepts as opposed to discrete, concrete sensorimotor abilities as opposed to abstract mental objects.

Concepts as...	**Concepts as...**
things we may reflect upon	things we possess and employ non-reflectively (Section 2.7)
intentionally imposed "top down"	activity derived "bottom up"
product of rational thought (rationalism)	product of empirical discovery (concept empiricism)
objects of perception	means of perceiving objects
consciously accessible	partly or substantially not (Section 3.3.1) consciously accessible
knowledge that	knowledge how (Section 2.7)
symbolic entities (sections 2.4, 2.6.1)	skillful abilities (Section 2.5)
"mental" representations (sections 2.4, 2.6.2, 2.6.4)	abilities to form representations (sections 2.5, 5.1)
sub-propositional components of thought	subconscious components of interaction
abstract and "mental"	concrete and "physical" (Section 4.1.2, 4.2.1)
abstracted from context	sensitive to context
"internal" to agent	"external" to agent – in environment (Section 1.5)
static	dynamic (Section 3.2.4)
discrete (individuable) (sections 2.1, 2.6.1.2)	continuous (sections 2.6.1.2, 5.2.4)
easily tied to language	clearly distinct from language (Section 3.3.2)
private entities	public entities (Section 3.3.3)

Table 5.1: Two contrasting views on concepts.

Table 5.1 summarizes each view in a set of descriptions with, as appropriate, links to the earlier sections. (If a link is given on the right side only, it applies to both columns.)

5.5 Conclusions

My conclusion in this chapter – that both concepts and conceptual understanding are bounded, that on pain of paradox, conceptual thought can never fully capture its own nature – sets me up against some pretty powerful opponents: people who would say that such a position equates to shrouding the mind in mystery, that providing a theory of human concepts or consciousness is no different from any other scientific endeavour. In many ways I would agree with this last point; but the lesson I think we should learn is not the one often taken, that science succeeds precisely by removing the observer, that a science of concepts or of consciousness will succeed by pressing a "pure" objectivity to the one place it has most resisted going. Rather, sciences of concepts and of consciousness remind us of what we should have remembered all along: that the observer is *always*

there (see again Section 2.8), that the subjective is inseparably bound up with the objective, that science yields up not timeless understandings freed from cultural and historical contexts but working hypotheses. It is not only the mind that we cannot know fully, but the world. Truth is a work in progress.

Chapter Five has attempted to do two important things: first, to draw lessons from the most important discussions of the preceding chapters; second, to lay the groundwork for Section 5.4 and the table I present there. If one could offer a slogan for this chapter, it would be that "concepts are necessary fictions (and a theory of concepts is, too)!" Both concepts and the theories about them are useful approximations that should not be mistaken for the things they purport to be about.

Theories of consciousness are currently more in vogue than theories of concepts. I offer reasons, in presenting the *hard problem of concepts*, to think that their fates are tied closely together. Both arise out of our pre-conscious experience and, at the same time, structure that experience, presenting a circularity that defies traditional models of causality – a challenge I will take up in Section 7.1. Echoing Chalmer's challenge for theories of consciousness, I argue that theories of concepts *must* take subjective experience as foundational, *and with that* a role for reflective thought and a place for representation.

Particular concepts may or may not be, implicitly or explicitly, self-referential. A theory of concepts, on the other hand, is unavoidably if implicitly self-referential: explaining concepts *as* a conceptual agent just is a circular enterprise; a theory of concepts is always put forward from within a pre-existing conceptual structure that assumes an experiencing agent. Conceptual understanding – including of concepts themselves – is always from some experiential point of view. Push the self-reference too far – try too hard to get at the "real" nature of concepts – and one is caught between vicious and merely pernicious circularities and by the known human limitations in disentangling recursive structures.

To conceptualize is to *step back* from the present moment and context, to stretch out that moment and consider it in light of previous ones and potential *what-might-be* moments to come. In the process, strict experience-in-the-moment is lost; a conceptual barrier is constructed between self and world. The conceptual illusion is that there is no barrier, no distorting simplification.

Paradoxes – my attention is mainly on explicitly self-referential ones – arise whenever we press up against our conceptual boundaries. I examine Roger Penrose's well-known argument that human understanding is *not* bound by Gödel's Incompleteness Theorem (a translation of self-referential paradox into the domain of formal mathematics and logic). I share with Penrose the understanding that Gödel's Theorem has something deeply profound to tell us about the nature of human cognition and understanding, even as I disagree over what lessons we should take away.

Finally, as the positive thesis to the prevailing negative thesis of the chapter, I restate an idea introduced in Chapter Two and developed in the chapters since: what I call the "toggling effect". I suggest that a theory of concepts, to truly address the nature of concepts as well as the debates in the literature, cannot settle on one single ontology or perspective, but will continually shift back and forth between two, both logically necessary and both logically incomplete, ultimately complementary, perspectives.

Chapter 6

Extending Conceptual Spaces

> If objects are represented as points in a conceptual space, then, roughly speaking, the similarity of two objects can be defined via the distance between the points that represent them in the space. Even abstract qualities may have a meaningful notion of distance (Aisbett and Gibbon, 2001, p. 196).
>
> When are two objects more alike to each other than each is to a third? Context and subjectivity have to be taken into account when answering this question (Aisbett and Gibbon, 1994, p. 153).

In the preceding chapters, I have attempted to work out the basic nature of concepts, offer a set of core properties, and put both concepts and the theories about them into an appropriate context. In this and the next chapter, I will assemble these various pieces together into a theory of concepts of my own, one that is firmly grounded in Gärdenfors' conceptual spaces theory of concepts. The lesson of the last chapter should be kept well in mind: if there are real limits to conceptual understanding, then any theory that attempts to understand that understanding will be similarly limited. The unified conceptual space theory, presented in the second half of this chapter, should be taken not as the correct theory of concepts – if I am right, no such single theory can exist – but rather my candidate for a "best answer to date", one I intend to describe clearly enough to then apply it to an account of concept acquisition and application (Chapter Seven) and to implement it in a simple computer program (Chapter Eight).

Rather than attacking other theories for what they get wrong, I would rather focus on what my approach gets right. On both of these points, I take inspiration from Gärdenfors.

I will begin the chapter by summarizing Gärdenfors' account, which is critically dependent on a notion of similarity. Indeed, conceptual spaces theory is an example of a *similarity space theory* (as noted in Section 2.4) – which, on the face of things, puts it in company with prototype and exemplar theories of concepts, in the empirical tradition, and in contrast with e.g. theory theory and informational atomism, in the rationalist tradition. That, however, is at best an over-simplification, and indeed, as is clear from passages in his book, Gärdenfors sees his theory as being compatible with, and complementary to, these other approaches. Consider for example what he writes about theory theory:

> . . . I believe that most of the role that the theories are supposed to play in representations of concepts can be taken over by the dimensions or domains that are considered

> to have the highest salience. In many cases these domains may not be perceptual, but indeed correspond to what is conceived of as *theoretical entities* in science.... From this perspective, there may be no fundamental conflict between a similarity-based theory of the kind developed in this book and a theory-theory (2004, p. 109).

Gärdenfors offers his theory as a bridging account between different levels of explanation of cognition more generally, and different accounts of concepts more specifically. I believe his theory is well placed to support the "toggling effect" thesis I proposed in Section 5.4.

Similarity-space-based theories are often criticized, and rightly so, for relying on a problematic notion of similarity. On these accounts, concepts that are somehow intrinsically more similar are grouped closer together, while those that are more dissimilar are grouped further apart. Fodor has argued (1998, p. 32), quite convincingly I think, that such an approach is doomed to failure, for invariably the measures of similarity that are being assumed depend upon an underlying layer of strict identity. For example, if two prototypes are more similar or less similar depending on how many features they share, then those features must, on pain of eternally receding target, be strictly identical[1]. Even more basic is the way that, as Janet Aisbett and Greg Gibbon note, similarity judgments are critically dependent on context (or, as Prinz describes it, "any two objects resemble each other in one way or another" (Prinz, 2004, p. 31).

The problem, as I discussed in Section 2.3.2 in the context of Goodman's generally-taken-as-devastating attack on resemblance, is not similarity *per se* but its position in the order of explanation. Gärdenfors avoids this trap, I think, by turning the order of explanation around so that, at the most basic level, things are similar because they are clustered close together in a common conceptual space.

Of course, there is an important sense in which similarity and similarity judgments are neither explanations nor explananda but rather to be taken as fundamental (at the same time as being highly context dependent). That comes down to our judgment of two things being the same or different with respect to one or more of their integral dimensions, and to the minimal perceptually or conceptually distinguishable unit of difference with respect to those dimensions[2]. By way of the latter, we can say of three things taken to share one or more integral dimensions, which two are most similar and which two are least. Furthermore, we can measure the similarity of any two things located within a common conceptual space as the distance between them according to whatever metric defines that space (Gärdenfors, 2004, pp. 5, 110)[3][4]. We can do all of this without falling afoul of Goodman's dictum.

Such similarity as I am describing here is characteristically interpreted as perceptual similarity, directly grounded in sensorimotor engagements with one's environment. The more one presses the

[1] That said, I think Fodor fails to acknowledge that if similarity is problematic, then so is identity, for much the same reasons. (See Section 2.3.2.)

[2] For the difference between the minimally perceptually distinguishable and the minimally conceptually distinguishable unit, in the context of the "fineness of grain" argument for non-conceptual content, see e.g. (Peacocke, 1992).

[3] This is, I think, the correct way to interpret the phenomenon of *categorical perception*, discussed by Gärdenfors in (2001, p. 387): "When categorical perception is at work, stimuli related to a specific category are perceived as indistinguishable, whereas stimuli from a 'nearby' category are perceived to be entirely different.... In color perception, for example, different shades of green are perceived to be more similar than green and yellow even though the wavelength differences are no larger."

[4] Aisbett and Gibbon (1994; 2001) have made a mathematically well-founded start toward defining such metrics formally.

explanation toward lower-level cognition, the more prominent a role this sort of similarity plays; the more one presses it toward higher, more abstract and symbolic levels, the less prominent the role of perceptual similarity and the greater the role of metaphor, where:

> The core hypothesis... is that *a metaphor expresses an identity in topological or geometrical structure between different domains.* A word that represents a particular structure in one domain can be used as a metaphor to express the same structure in another domain (Gärdenfors, 2004, p. 176).

Metaphors[5] are often used as a way of making abstract ideas concrete. Abstract things are typically couched in metaphorically sensorimotor-based terms: e.g., "cultivating hope" (planting, watering, etc.). What matters for present purposes is that, at either level, one has similarity, as a continuous relation, being grounded in something discrete: on the one hand, a metric distance; on the other, an identity of structure.

This chapter will set out what I take to be the essential theses of conceptual spaces theory, putting it in the context of historical and contemporary theories of concepts, as well as relating it to other trends within philosophy of mind: specifically, the enactivist tradition associated most strongly with Francesco Varela. I will evaluate the empirical testing of the theory to date and consider its current limitations.

In the second half of the chapter I will lay out my plan for moving from many conceptual spaces to a single unified conceptual space – one that is more directly algorithmically describable[6]; using a set of building blocks and rules for joining them together or taking them apart: the equivalent, metaphorically, of describing the playing board, providing the pieces, and specifying the rules of a game. By focusing on the algorithmically describable, such an approach comes not only with strengths (e.g., easier to implement in a computer model, easier to be clear what is being tested with that model) but important limitations (downplaying that which is not easily reduced to algorithms); but the limitations, properly understood, can themselves, I will argue, provide a source of strength.

6.1 Conceptual Spaces Theory

> The fundamental cognitive role of concepts is to serve as a bridge between perceptions and actions (Gärdenfors, 2004, p. 122).
>
> ...I argue that conceptual spaces present an excellent framework for "reifying" the invariances in our perceptions that correspond to assigning properties to the perceived objects (Gärdenfors, 2004, p. 59).

To reify is, of course, to give concrete expression to something abstract. In this way concepts play a role similar to metaphor; in an important way – thinking here of Section 5.2.3 – concepts *are* metaphors (see the fifth point below). One of the most striking aspects of concepts is the way they bring together the very abstract and the very concrete: showing how they relate, shifting the focus between the two constantly. One might be tempted to go further than Gärdenfors and suggest that conceptual spaces theory presents an excellent framework not only for understanding (within

[5] Note that Gärdenfors takes metaphor to be primarily a semantic not a linguistic issue.
[6] For an alternate, more formal approach with much the same intentions, see (Aisbett and Gibbon, 2001).

the limits of our conceptual abilities) the nature of the invariances he discusses, but also for telling a compelling story both of how they arise out of our perceptions and how, at the same time, they structure the very perceptions that give rise to them. This will be the theme of Chapter Seven.

The central tenets of conceptual spaces theory I take to be:

- Neither associationist (including connectionist) nor symbolic accounts of cognition, and likewise neither empiricist nor rationalist approaches, can, on their own, do adequate justice to the nature of concepts. The former are too reductionist, the latter too rarefied. Associationist and symbolic accounts should be understood as two different levels of explanation of cognition, which a theory of concepts should then try to bridge.

- Just as a conceptual account of cognition bridges these two levels of explanation of cognition, concepts themselves bridge two levels of cognition. That is, they sit in the middle between directly sensorimotor-grounded cognition on the one hand, symbolically and propositionally structured thought on the other. The former is more unconscious and automatic, and shades over into the subpersonal; the latter is indisputably personal, and much if not most of the time conscious and deliberate[7]. Concepts are beholden to neither one level nor the other.

- "There is no unique correct way of describing cognition" (Gärdenfors, 2004, p. 2). "In brief, depending on which cognitive process we are trying to explain, we must choose the appropriate explanatory level" (Gärdenfors, 2004, p. 57). Likewise, there is no unique correct perspective on concepts. Depending on which aspects of them we are trying to explain, they may look more like words of a language, say, or more like patterns of association.

- Furthermore, there is no unique correct perspective on any particular concept, not least because concepts *change*: with the agent who is using them and with the context in which they are used (see Section 3.2.4). Gärdenfors specifically includes the so-called *natural kinds* concepts (see Section 3.4.3.1).

- A metaphor for physical objects in physical space, concepts are (best understood as) either:

    1. *points* (or associated sets of points) within conceptual spaces, whose dimensions (e.g., hue, saturation, and brightness in the case of colour) may be acquired in a bottom-up activity-driven manner or a top-down intentionally-driven one[8]; or as
    2. *shapes* (or associated sets of shapes) within those same spaces.

- The upshot of this is, as I read it, that there is no principled class/instance distinction to be made (*cf.* the similar point made in Section 2.1): any particular instance of a concept (a point) can, within practical limits, be expanded to a shape (a "class" or set of points, each one a *more particular* instance), and any shape can be collapsed to a point (i.e., treated as an instance of some, more general, concept)[9].

[7] The personal/subpersonal distinction was first raised by Dennett (1969), where, for him, the personal is, roughly, "for an (entire) agent", while the subpersonal is, roughly, everything below that: the nervous system events and so on that make the agent possible.

[8] Compare: "In a conceptual space that is used as a framework for a scientific theory or for construction of an artificial cognitive system, the geometrical or topological structures of the dimensions are *chosen* by the scientist proposing the theory or the constructor building the system.... In contrast, the dimensions of a *phenomenal* conceptual space are not obtainable immediately from the perceptions or actions of the subjects, but have to be *infered* from their behavior" (Gärdenfors, 2004, p. 21).

[9] So for example, **WEIGHT** is both a specific instance of **MEASUREMENT** and itself a class of different weight values: e.g., 13 kilograms. **THIRTEEN KILOGRAMS** is both a specific instance of **WEIGHT** and a class of weight values between (typically) 12.50 and 13.49 kilograms.

- Those shapes are typically (though not always) *convex* shapes: that is, for any two points that lie within the concept x, all points on a straight line between them should also lie within that concept. (Some concepts are defined as the negation of other concepts, within a certain domain: e.g., Gentiles are anyone who is not Jewish. Fodor uses the example of **NOT A DUCK**. If the one concept [**JEWISH** or **DUCK**] is convex, its negation [**GENTILE**, **NOT A DUCK**] within a domain cannot be.)

- Individual convex shapes (or individual points) denote a particular type of concepts, namely properties – what I have preferred to call (see Section 4.2.2.3) *property concepts*. Property concepts relate only to a single domain: e.g., **COLOUR** can *only* be described in terms of the integral dimensions **HUE**, **SATURATION**, and **BRIGHTNESS** (or their equivalents), and never according to some entirely different set of integral dimensions such as **MASS** and **DENSITY**. Other types of concepts (what I have referred to as *object concepts* or *action/event concepts*) are associated sets of these shapes (or points) across multiple domains. Note that, just as individual shapes can be collapsed to points, so, too, associated sets of these shapes can be collapsed to a single shape: i.e., *all* concepts can be treated as property concepts. This is not explicitly stated, but I take to be implicit in Gärdenfors' account[10].

- The structure of concepts need not in any way be consciously introspectible: "...For many words in natural languages that denote properties, we have only vague ideas, if any at all, about what are the underlying conceptual dimensions and their geometrical structure" (Gärdenfors, 2004, p. 168).

- The process of "carving up" a conceptual space into various shapes and sub-domains is the process of *categorization*: "...Where (possible) objects are represented as points in conceptual spaces, a categorization will generate a partitioning of the space and a concept will correspond to a region (or set of regions from separable domains) of the space" (Gärdenfors, 2004, p. 60).

- At the same time, that "carving up" imposes a *Voronoi tessellation* on the space (see figures 6.1 and 6.2). A Voronoi tessellation tiles a plane that is initially populated by a set of points (the Voronoi *sites*), which in conceptual spaces theory are taken to represent the most prototypical members of a category. The plane is then divided up according to which of those points the remaining points in the plane are closest to (the Voronoi *cells*). Boundaries arise wherever there is equidistance to two of the existing points, junctions wherever there is equidistance to three (or more) points. Although these tessellations are in two dimensions, a three- or n-dimensional Voronoi tessellation can also be done[11].

Likewise, **CAT** is a specific instance of **MAMMAL** and itself a class of various breeds of cats. **MY CAT KALI** is both a particular cat and a class of all my experiences of Kali, in different times and contexts. At some point, the instances-to-classes expansion has to bottom out, for strictly practical reasons: **MY CAT KALI AT 12:03 PM YESTERDAY** may be as specific as my sensory experience and memory allow.

[10] By analogy, think of the way that, in English, in general, all nouns can take the role of adjectives: e.g., "**bicycle** thief", "**political correctness (PC)** police". Verbs do something similar but change their form: "**cycling** champion".

[11] Aisbett and Gibbon express this more formally:"Under its usual definition of convexity, any Euclidean space can be divided into a set of n convex regions, disjoint except at boundaries, where the ith region is chosen to contain a pre-specified point x_i. This is done in a process called Voronoi Tessellation, which assigns points in the space to the ith region if and only if x_i is the closest of the prespecified points..." (Aisbett and Gibbon, 2001, p. 200).

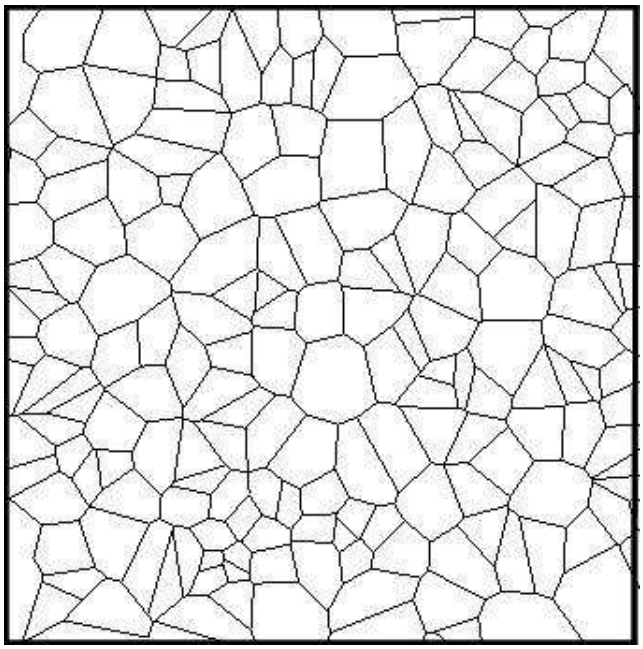

Figure 6.1: A randomly generated Voronoi tessellation.

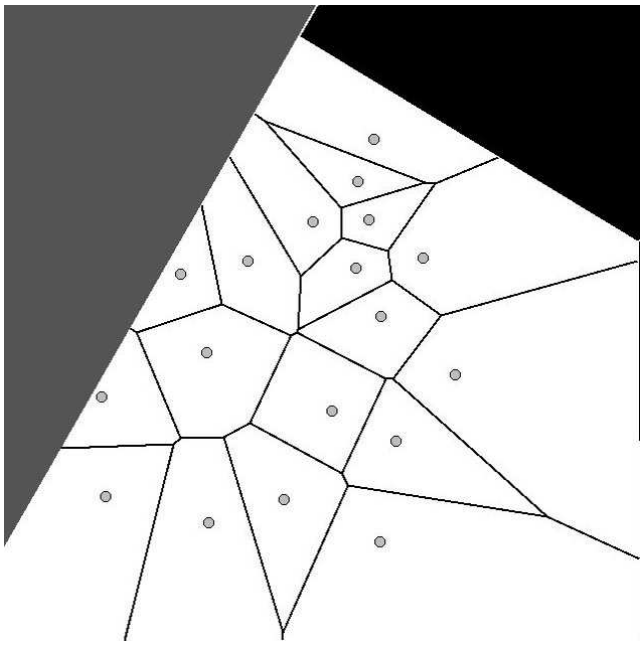

Figure 6.2: A non-random tessellation created by the program described in Chapter Eight. Unlike the first picture, here the initial points (the Voronoi sites) are indicated within each cell.

6.1.1 Comparison to Other Theories of Concepts

> As we have seen, the delineation of natural properties in terms of convex regions... provides intuitively plausible solutions to the problems caused by the account within intensional semantics. Apart from this, the criterion derives independent support from the *prototype theory* of categorization developed by Rosch and her collaborators.... The main idea of prototype theory is that within a category of objects, like those instantiating a property or a concept, certain members are judged to be more *representative* of the category than others (Gärdenfors, 2004, p. 84).

Conceptual spaces theory has its historical roots in the imagist tradition, whose modern heirs include prototype and similarity space theories (see Section 2.4). Among contemporary accounts, it seems to fit most comfortably with the mental representations rather than the abilities camp, but this may be only because of Gärdenfors quite frequent, and to my mind somewhat unfortunate, references to representations, to which he appears to apply a very broad definition. That Gärdenfors considers himself a representationalist is not in doubt, nor is there any doubt that he is not an anti-representationalist, as that term is generally used. What is much less clear is how much he has in common with other self-avowed representationalists (I have in mind someone like Fodor), given that he is equally happy talking about concepts as abilities. Also, representationalism is frequently associated with indirect realism, and Gärdenfors, as previously noted, is an anti-realist.

In a welcome break from the general standard of writing in philosophy of mind, Gärdenfors' objective is not to attack his competitors' theories but rather to focus on his own, positive thesis. Nonetheless, he does locate himself relative to various alternative theories of concepts and, in the process, make clear how his approach differs from each of them. In particular, he takes care to note his debt to prototype theory (e.g., "learning a concept often proceeds by generalizing from a limited number of *exemplars* of the concept.... Adopting the idea that concepts have prototypes, we can assume that a typical instance of the concept is extracted from these exemplars" (2004, p. 123))[12] and show how the two approaches support each other ("...if prototype theory is adopted, then the representation of properties as convex regions is to be expected, at least in metric spaces." (2004, p. 87)). Nonetheless, Gärdenfors is not entirely uncritical of prototype theory, noting its difficulty accounting for concept acquisition (2004, p. 123) as well as its inability to address the size of any particular category ("ducks" is arguably a broader category than "ostriches") or the degree of variability among its members (i.e., how tolerant the category is to non-typical examples) (2004, p. 138).

Gärdenfors does not address Prinz's "proxytypes" variation on prototypes, though I would expect him to be critical of proxytypes' somewhat hodgepodge structure, in contrast to conceptual spaces theory's own relative descriptive uniformity: to wit, "a proxytype can be a detailed multimodal representation, a single visual model, or even a mental representation of a word (e.g., an auditory image of the word 'dog')" (Prinz, 2004, p. 149). I think he would be critical as well of proxytypes' modal dependence, in contrast to conceptual spaces theory's own amodality. (I use the term "amodal" here guardedly. Neither Gärdenfors nor I would wish to imply that concepts "really are" independent of sensory modalities, only that they can be given a modally independent description.)

Likewise, Gärdenfors does not directly address Fodor's informational atomism, though he does critique Fodor's language of thought (LOT) hypothesis, which is kith and kin with informational atomism, and which he criticizes for its inability to address the symbol grounding problem (see Section 2.6.1.1). Conceptual spaces theory leaves room for a version of LOT, but it is one that Fodor, with his insistence on the functional independence of the conceptual level of cognition from the underlying levels implementing it, surely could not endorse.

Gärdenfors reserves perhaps his clearest, though still mild, criticism for theory theory, which he criticizes, as I wish to as well, for what seems to be an inescapable vagueness:

[12] Note that in contrast to typical exemplar theories, neither prototype theory nor conceptual spaces theory assume that the prototype need be among the exemplars.

> The basic idea [of theory theory] is that concepts should be thought of as embedded in knowledge that contains *theories* of the world.... One weakness of theory-theories, however, is that they hardly give any account of *how* the assumed theories are represented in cognitive systems. And, assuming they are expressed symbolically, this would amount to putting the cart before the horse.... Furthermore, it is not clear *what* can count as a theory; if everything can, then the theory-theory is empty (2004, pp. 107-108).

Bottom line: the concept of theory is so abstract and can be applied so broadly as to make it of limited usefulness to a theory of concepts.

6.1.2 As Located Within an Enactivist Framework

> I have proposed using the term *enactive* to... evoke the idea that what is known is brought forth, in contraposition to the more classical views of either cognitivism or connectionism (Maturana and Varela, 1992, p. 255).
>
> *Everything that is said, is said by an observer to another observer that could be himself* (Maturana, 1978, p. 30).
>
> The roots of mental life lie not simply in the brain, but ramify through the body and environment. Our mental lives involve our body and the world beyond the surface membrane of our organism, and so cannot be reduced simply to brain processes inside the head (Thompson, 2007, p. ix).

Although not explicitly enactivist, it is easy to locate conceptual spaces theory within an enactivist framework, as Gärdenfors himself acknowledges[13]. Like Varela, Gärdenfors wishes to contrast his approach with the "classical views" of cognitivism and connectionism, and to offer a view of knowledge as something that is never static but always in the process of being "brought forth" through the interactions of agent and environment, where neither can be cleanly separated from the other.

Despite its implicit reference in the title ("Concepts Enacted"), enactivism[14] has so far been lurking in the background. In this chapter and the next it will take a more critical role. For now, it will suffice to say more precisely what I mean by the term and relate it to Gärdenfors' work.

Enactivism, as I wish to use the term, should not be confused or equated with the twin notions of embeddedness and embodiment, as much as it does embrace them. Embeddedness (or *situatedness*) is the way an agent is located in a particular spatio-temporal context. Embodiment is the way an agent takes a particular physical form, which constrains its interactions with its environment[15]. Enactivism goes beyond embeddedness/embodiment by:

- Understanding cognition, at least in the first instance, as a *skillful activity*, and in any case as a lived, dynamic process and not a static entity[16].

[13] Personal communication.
[14] One will frequently see the same ideas under the label of "enaction".
[15] Compare Gärdenfors: "Conceptual structures are *embodied* (meaning is not independent of perception or of bodily experience)" (2004, p. 160).
[16] This was the theme of the June 2010 Enaction Summer School in County Tipperary, Ireland: http://www.enactionschool.com.

- Typically perceiving continuities as underlying that which appears individuable and discrete, most notably, the continuity between agent and environment, such that the agent's mental life extends, in a meaningful way, into the world (see the opening quote from Thompson, and compare the discussion in Section 5.2.3.2 about the extended mind hypothesis).

- Taking an agent/environment, internal/external distinction to be both conceptually necessary and, at the same time, meaningful *only with respect to an observer*, and not to the organism itself independently of some identifiable observer (see the opening quote from Maturana, which could be taken as a defining statement for *constructivism*).

- Giving a foundational role to phenomenology and emphasizing the essential contribution to be made by first-person methods (see e.g. (Varela and Shear, 1999)).

Enactivism is something of a term of fashion at the moment. As one of the people evaluating abstracts for a major consciousness conference told me in a personal communication, well over a third of the abstracts read by that reviewer for the conference included the word in their title, the reviewer counting its inclusion, in general, a count against the abstract.

The problem is, as with most terms of fashion, that the word means widely different things to different people. Contemporary usage has much of its roots in a book (***The Tree of Knowledge***) by Humberto Maturana and Francesco Varela, for whom, in many ways, the concept is bound up with another notion, *autopoiesis*. Autopoiesis is intended as an alternative description of what qualifies as a living organism, in terms of operational closure (processes of the system are produced from within the system; anything external to the system can play only the role of catalyst), autonomy, and the observation that organisms "are continually self-producing" (Maturana and Varela, 1992, p. 43). On the other hand, Noë has called his own approach to cognition "enactive" but does not talk about autopoiesis, is more specifically focused on sensorimotor engagements (and less on the coupling between cognition and life), and is more recognizably (and self-avowedly) externalist (rather than seeking to avoid either internalist or externalist labels).

Conceptual spaces theory, as well as my own position, fits in best, I think, with the enactivism of Maturana and Varela, or such contemporary philosophers as John Stewart or Evan Thompson, both of whom published with Varela. There are, however, certain caveats[17]:

- Conceptual spaces theory lacks their often strongly anti-representational bias and is even favourably disposed toward representations, properly understood[18].

- In consequence, conceptual spaces theory is favourable to their view of cognition as skillful activity, but only when interpreted sufficiently broadly as to leave room for representations in a way that bridges the knowing that / knowing how divide (see Section 2.7).

- As with Noë's project, conceptual spaces theory recognizes the relationship between cognition and life but is not concerned to make such a tight coupling between them. (Its focus is

[17] All of these will, again, describe my own position.

[18] How much this is a terminological dispute is difficult to say, given how frequently representations are discussed without being defined, on both sides of the representational/anti-representational divide. Clearly some usages of "representation" are so broad as to be practically vacuous. I believe that enactivism makes a mistake here, allowing dogmatism to get in the way of its more deeply held principles that would seem to *require* a place for representations within the (ineliminable) observer perspective: representing is what observers *do*. That implies that where there is no observer, there will be no representations. But enactivists should be more aware than most people that there is no way of getting through to the "real" (independent of any observer) fact of the matter.

not so much on self-organizing systems, and it makes no mention of anything resembling autopoiesis.)

- Conceptual spaces theory is therefore more sympathetic to the possibility, at least, of artificial intelligence as distinct from artificial life.

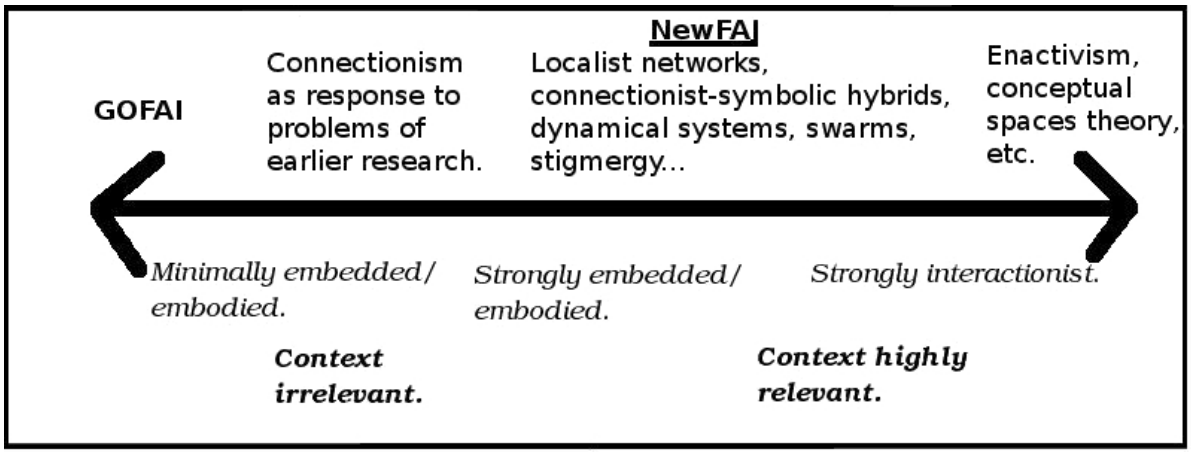

Figure 6.3: From GOFAI (Good Old-Fashioned AI) to "NewFAI" and beyond.

As Figure 6.3 suggests, both enactivism and conceptual spaces theory can be seen in the context of the history of cognitive science as part of a broader movement away from largely disembodied and "purely" symbolic accounts of cognition that treated e.g. agent as independent from environment, sensory input as independent from motor output, mind (software) as independent from brain (hardware), cognition as independent from life, syntax as independent from semantics, and so on[19]. At the same time, neither should be taken as final destinations (conceptual spaces theory is quite clear about this) but only as points along a path.

6.1.3 Empirical Testing to Date

> My ambition here is to present a coherent research program that others will find attractive and use as a basis for more detailed investigations (Gärdenfors, 2004, p. ix).

One might well rue the frequent disconnect between abstract theory and empirical testing, and nowhere may this be clearer than in the intersection between philosophy of mind and cognitive science. The analytically inclined philosophers of mind decry the continental philosophers for their lack of empirical grounding, and yet their attempts at naturalization have met, at best, with mixed results. Theories translate imperfectly into implementable models, and empirical results are nearly always open to interpretation. It can be difficult to find the middle ground between armchair reflection on the one hand and applications of dubious theoretical import on the other, and it might seem that there is a tendency to slide off in one or the other direction.

One philosopher who has taken particular pains to exploit the middle ground is Ron Chrisley with his work on expectation-based architectures[20] and the SEER-3 robotic platform (Chrisley and

[19] Of course choices were (and are) made based on the computing resources and the supporting theories available at the time. I do not wish to caricature "good old-fashioned AI" (GOFAI), within which framework a lot of good research was done (see e.g. (Whitby, 2003)).

[20] Roughly, this is the role expectations play in being partly constitutive of experience: i.e., expectations as the *products* of experience also *shape* experience.

Parthemore, 2007a,b). The theory informs the robot model and the model informs revisions in the theory, which then informs revisions in the robotic model, with the goal of making the loop as tight as possible and the number of iterations as great as possible (more on this in Chapter Eight). The robot model in this way serves both as demonstration of the theory and intuition pump for the theorist.

What then of conceptual spaces theory? As with most of the present work, conceptual spaces theory is strongly on the theoretical side of the divide. At the same time, Gärdenfors has attempted to create a theoretical structure that *invites* testing.

Three papers deserve mention here. The first (Gärdenfors and Williams, 2001) does not present any new empirical research but rather seeks to locate conceptual spaces theory in the context of existing evidence in psychology for prototypes as they relate to categorization, and indeed argue that conceptual spaces theory can provide a *better* account of that relationship, one that relies on computation rather than fuzzy intuition. Conceptual spaces theory is given a more algorithmically precise formulation by relating it to something called the Region Connection Calculus (RCC). More algorithmically precise is, of course, easier to implement in a computer model (or test in a psychology experiment) – the very goal of the unified conceptual space theory presented in the second half of this chapter. Voronoi tessellations are used to determine category boundaries, and then the RCC is used to reason about them. Particular attention is paid to the "crisping" or "blurring" of boundaries, and how that may be used to account for non-monotonic reasoning (i.e., "if X then Y, *ceteris paribus*").

The second (Chella et al., 2004) *does* offer new empirical research – involving two mobile robots, each using conceptual spaces to navigate their environment – but the account (less than half a page out of a six-page paper) is extremely brief, making it difficult to know what has actually been implemented and what conclusions can reasonably be drawn. (It should be noted, too, that issues of salience for these robot models have been pre-determined and object recognition hard-wired: i.e., their "concepts" of physical objects are not acquired by the robots but rather provided by the researchers.)

The main concern of the paper is less to present empirical results than, again, to lay the groundwork for further empirical testing: in this case, focused on what the authors call *perceptual anchoring*, which they take to be a special case of Harnad's (1990d) symbol grounding problem. They define perceptual anchoring as "the problem of creating and maintaining in time the connection between symbols and sensor data that refer to the same physical objects" (Chella et al., 2004, p. 40). Conceptual spaces again play a bridging role by allowing one "to represent discrete concepts, which are the main entities manipulated at the symbol level, inside a structure where we can place continuous observable quantities, which are the main entities provided by the perceptual system" (2004, p. 40). The pseudocode procedures for finding, tracking, and acquiring (or re-acquiring) anchors relate nicely to my account in Section 6.2 of examining one's unified conceptual space in terms of the queries "What is here?", "Is this here?", and "What if this were here?"

The final paper (Chella et al., 2008) extends the ideas about perceptual anchoring. It offers the best glimpse into how conceptual spaces theory might be tested empirically and applied concretely, in this case within the emerging (and still quite controversial!) field of machine consciousness.

Here, much of the emphasis is on meta-cognition: "We claim that one of the sources of self-consciousness are *higher order* perceptions of a self-reflective agent" (2008, p. 153) (see Section 4.1). Unfortunately, for present purposes, much of the (again short) paper is devoted to describing the robot's "conceptual area" (one of three cognitive levels being modeled, the other two being the "sub-conceptual area" and the "linguistic area") in terms of low-level mathematical justifications rather than high-level algorithmic descriptions. So again, it is hard to know, with respect to the present discussion, what precisely has been implemented or tested: notably, the question of how the robot is to deal with the sheer volume of possibilities opened up by having first-, second-, and higher-order concepts all available within its conceptual area is touched upon but not resolved. However – unlike in the earlier robotic system – salience (the so-called *frame problem*) *is* addressed in a preliminary way, and indeed given a treatment quite reminiscent of Chrisley's work on expectation-based architectures.

6.1.4 Limitations and Difficulties

> Philosophers will complain that my arguments are weak; psychologists will point to a wealth of evidence about concept formation that I have not accounted for; linguistics [sic] will indict me for glossing over the intricacies of language in my analysis of semantics; and computer scientists will ridicule me for not developing algorithms for the various processes that I describe. I plead guilty to all four charges (Gärdenfors, 2004, p. ix).

I have said earlier that I find Gärdenfors' use of the term "representation" over-broad, including much that I would prefer not to call representational. I have attempted my own account of what should and should not be called a representation (see Section 2.6.2). But this is a minor point.

The far greater challenge to conceptual spaces theory is filling in the details. In contrast to much of the literature in this area, Gärdenfors is refreshingly modest and candid about how, in many ways, his theory provides only the scaffolding – and like all true scaffolding, it is meant to be removed once the structure (which might itself be the scaffolding for yet *another* structure) is in place. Although he offers a few different examples, Gärdenfors focuses much of his attention on the concept of colour, with its relatively uncontroversial dimensions of hue, saturation, and brightness. For many if not most other concepts, it is unclear how one is meant methodically to go about determining their integral dimensions.

Two missing details are of particular importance and seem to me to capture many of the others one might name. First is the way an agent's many different concepts, for all that they are quite divergently structured, nevertheless have a common underlying structure or blueprint; they are all meant to be built up out of primitive elements in more or less the same way, by the same set of (implicit) rules. Their integral dimensions should be determined in more or less the same manner.

Second, and closely in line with this, is the way that an agent's many conceptual spaces come together in a single unified space of spaces. For it was never meant to be the case that one's **COLOUR** concept exists in one conceptual space, one's **SHAPE** concept in another, without their being, at the same time and at a sufficient level of abstraction, part of a common space – likewise for any two concepts you can name.

The uniform conceptual space theory that occupies the second half of this chapter is meant to address both of these shortcomings. It is meant, as well, to address the fourth of Gärdenfors' four "charges" above. In particular, when it comes to talk of concepts as the building blocks of thought, I hope to put some more literal flesh onto the bones of that familiar metaphor.

6.1.4.1 Uniform Concepts

> If we assume that concepts are structured, a question quickly arises about whether any one kind of structure is suitable for all kinds of contents. . . . If concepts are structurally uniform (or uniformly unstructured), a uniform theory of concepts is easier to achieve (Prinz, 2004, p. 94).

As Machery has noted (see Section 5.2.1.2), most theories of concepts are trying to achieve something close to a uniform account of concepts. Prinz points out that such uniformity is easier to achieve if concepts themselves are uniformly structured. In the case of a conceptual-spaces-type theory, uniformly structured concepts are meant to populate uniformly structured conceptual spaces and, ultimately, give rise to one uniformly structured unified space[21].

6.1.4.2 Unified Space

> . . . The model is tailored to representing whole semantic fields: There is always a conceptual space projected by the whole range of possible feature values, and this space is then divided up into a certain number of conceptual categories. It does not make very much sense to project a whole such space for the representation of one single item. In other words, we can say that the model is inherently contrastive – in distinction to the symbolistic approach, in which features, and hence semantic categories, can exist "by themselves" (Geuder and Weisgerber, 2002, p. 71).

[21] "Uniform" should be read in all of these cases as "roughly uniform".

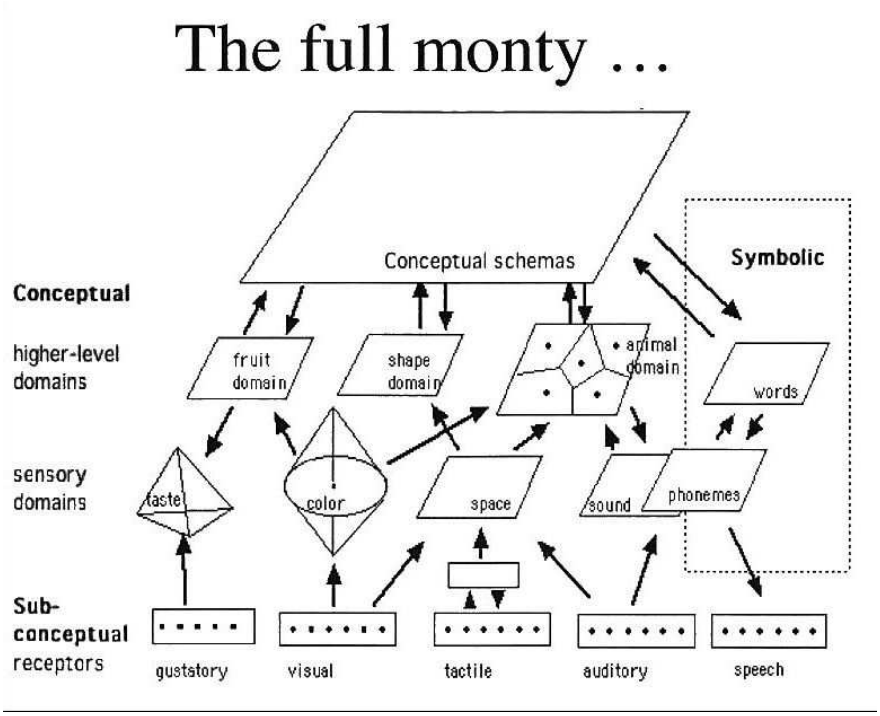

Figure 6.4: The Full Monty. Diagram by Peter Gärdenfors; reproduced by permission.

Once one allows that there are such things as concepts, one has an obligation to explain how those concepts form part of a single (roughly if not strictly) consistent system, that is complete in the sense that it incorporates *all* of an agent's concepts. Gärdenfors does offer some clues about how the different spaces are meant to map onto one another, both in his description of "natural concepts" being "represented as a set of regions in a number of domains together with an assignment of salience weights to the domains and information about how the regions in different domains are correlated" (2004, p. 105), and in his extended discussion of metaphor (2004, pp. 176-186). His original intention with ***Conceptual Spaces*** was to include a final chapter on creating a unified space, but he was not satisfied with what he had and left it out[22]. Figure 6.4 was intended for that "missing" chapter.

Wilhelm Geuder and Matthias Weisberger attempt to show how one might go about constructing a unified conceptual space in the subregion of action concepts, clearly intending, I think, for it to apply more widely. In doing so, they contrast the conceptual spaces approach with e.g. Fodor's informational atomism, in which Fodor is quite explicit that particular concepts could, in principle, exist in utter isolation from one another (see e.g. (Fodor, 2008, p. 54)). In conceptual spaces, concepts exist within a universe of concepts, which collectively they define.

6.2 The Unified Conceptual Space Theory

According to the uniform conceptual space theory, as put forward in the remainder of the chapter, all concepts exist within a common space that is the mapping together of many different conceptual spaces, all with the same general structure: a *space of spaces*[23]. Points within that space should be addressable – not necessarily in absolute terms, but certainly in relative ones (i.e., relative to

[22] Personal communication.
[23] An earlier version of many of these ideas is to be found in (Parthemore and Taylor, 1992).

another location in the space). What I mean is, one should be able to use the unified space (or a model of it) to answer questions about a particular location in that space, such as:

- *What is here?* Return the contents of that location.
- *Is this here?* Allow mappings or comparisons between different parts of the unified space. Given a location and an expectation of what to find there, say whether or not the expectation is met.
- *What if this were here?* Allow simulations and possible-worlds-type scenarios. Given a location in the unified space, allow substitution of what is *not* found there (in terms of what might be or what might have been) for what is.

I am speaking here in the first instance of concepts *for an individual*, though an analogous space must logically exist for a society – for those conceptual agents who are social animals – mapping together the different conceptual spaces of its members into a unified space of the whole. (I do not mean by this, in any way, to imply that the direction is necessarily from the individual to the society. Indeed, Pierre Steiner and John Stewart have offered a convincing argument (Steiner and Stewart, 2009) that much of social cognition does *not* begin with or reduce to an agglomeration of individuals, but must be taken as foundational to cognition. A similar view may be found in (Jaegher et al., 2010). Rosch, who has long argued that categories are intrinsically cultural artefacts – see e.g. (Rosch, 1999, p. 189) – would likewise be inclined this direction.)

A key insight of the unified conceptual space theory is that concepts can be given two contrasting structural descriptions: one from geometry, one from logic. The first (Section 6.2.1) is borrowed straight from conceptual spaces theory: concepts may be described as well-behaved and normally convex shapes within the unified space. *These shapes can, optionally, be assigned an arbitrary*[24] *symbol or label, that then accrues to all points within that subspace*[25]. The second (Section 6.2.2) is only hinted at in Gärdenfors' theory: concepts may *also* be described as a structured set of logical relations to other parts of the unified space. This is to say, concepts are defined *both* by the concepts with which they are contiguous, within the same domain; and by the concepts to which they are in one way or another associated, in adjacent or distal domains.

6.2.1 Dimensions of the Unified Space

> One important aspect of the organization of conceptual knowledge has to do with the hierarchical structure of concepts. . . . This hierarchical structure can be interpreted as facilitating our thinking about objects and entities. The facilitation arises out of the relations between levels of the hierarchy and information associated with the different levels. Knowing what a mammal is and that a bat is a mammal, allows us to distinguish it from birds and group it together with other animals that nurse their young (Hemeren, 2008, p. 24).

If the notion of well-behaved shapes can be taken straight over from conceptual spaces theory, then something must still be said about the dimensions of the unified space in which they are all to be located. After all, on Gärdenfors' account, the dimensions of the individual conceptual spaces are nothing more than the integral dimensions in that domain – hue, saturation, and brightness in the

[24] "Arbitrary" is intended in the sense of Section 2.6.1.2.

[25] I will follow the convention through the remainder of the chapter of highlighting certain key axioms of the unified conceptual space theory by putting them in italics.

case of the colour space; and coming up with integral dimensions that are common to all concepts would seem to render the notion vacuous.

The unified conceptual spaces theory proposes, instead, that the unified space be described along four axes.

6.2.1.1 Axis of Generalization

The first axis describes a hierarchy of concepts from maximally general to maximally specific. This is the axis Hemeren is referring to. It is also the axis along which one should place Rosch's (1975; 1999) distinction between superordinate-, basic-, and subordinate-level categories.

The axis is divergent in both directions. Many (though not all) concepts can be assigned to multiple superordinate concepts: e.g., **PERSON** can be an instance of **BIOLOGICAL ORGANISM** or **INTENTIONAL AGENT**. (On the other hand, **BROWN** is only an instance of **COLOUR** – more in Section 6.2.2.3.) Likewise, as noted in Section 6.1, all concepts can, within limits of practicality, give rise to more specific instances.

6.2.1.2 Axis of Alternatives

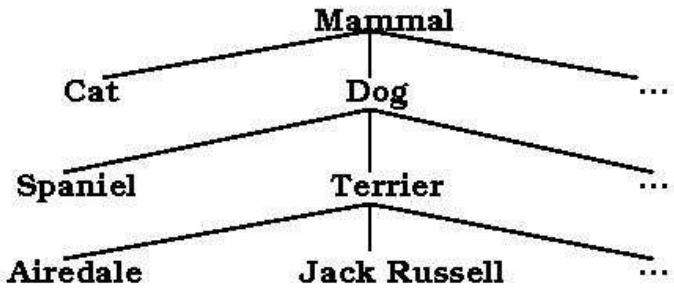

Figure 6.5: Diagram visualizing the first two of the three axes, reproduced with permission from (Hemeren, 2008, p. 25). Note that the labeling of the axes in the text is mine, not Hemeren's.

The second axis describes a range of alternative instances of a concept, at any particular level of generalization. These are arrived at by adjusting the value of one or more of the concept's integral dimensions and are arranged from maximally similar to maximally dissimilar along those dimensions (remembering the earlier discussion about similarity). This axis, also, is divergent in both directions, depending on which integral properties are being attended to (presuming there is more than one). Consider a shade of brown, which may be adjusted according to hue, saturation, or brightness, or some combination of the three. Compare this to moving around within the conical colour space. See Figure 6.5, which visualizes the axis of generalization (vertical) and the axis of alternatives (horizontal).

6.2.1.3 Axis of Dynamics

The third axis describes a range of concepts from most relatively static to most explicitly dynamic: from object-like to action-like things, from things described best by nouns to those described best by verbs. Depending on what is being emphasized, the same concept can be viewed along a certain range more toward one end of the axis or the other. For example, the concept of a human male

considered as an anatomical specification lies more toward the "static" end, whereas the concept of a human male considered as a living, acting, intending agent lies more toward the "dynamic" end.

6.2.1.4 Axis of Abstraction

The final axis describes a range of concepts from maximally non-conceptual ("zeroth order") to maximally higher-order-conceptual: i.e., from maximally concrete and physical to maximally abstract and "mental" (see Section 4.2.1). Note that, at its one extreme, it would converge with the axis of generality: a maximally general category and a maximally abstract one amount to the same thing. The converse is not the case, however: a maximally specific category (e.g., any acts of injustice that occurred at 14:03 hours yesterday at the corner of First Street and Seventh Avenue in New York City) need not be a maximally concrete/physical one.

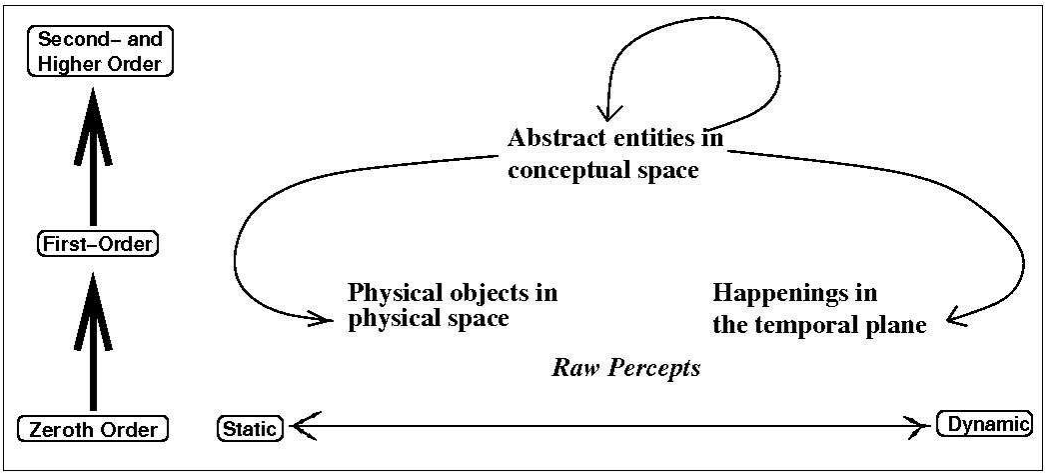

Figure 6.6: Diagram visualizing the third and fourth axes.

Figure 6.6 visualizes the axis of abstraction (vertical) and the axis of dynamics (horizontal). First-order concepts are concepts, roughly, of physical objects in physical space and of happenings in what one might call the *temporal plane*, whose one axis is a timeline from past to future and whose other axis is a succession of alternative *might be*s or *might have been*s[26]. As experienced, those objects and happenings are structured out of percepts that are themselves proto-conceptually structured and hence already interpreted (see the discussion in sections 5.2.3 and 7.2.3). Higher-order concepts are concepts of abstract entities in an analogous conceptual space. *Concepts, by their reference, map sub-regions of the conceptual space onto other sub-regions of the conceptual space, sub-regions of the physical space (relative to a moment in or span of time), or sub-regions of the temporal plane (relative to a point in or sub-region of physical space).*

The further along the axis of abstraction one goes, the more recognizably the entities one is dealing with are concepts and not non-concepts. Concepts are abstract entities in conceptual space, just as physical objects exist in physical space and happenings can be located on the temporal plane. As such, they may describe those physical objects or happenings, or they may describe other abstract entities like themselves: hence the arrows.

[26]One must distinguish here between experienced time, which has one dimension; and imagined time, which encompasses experienced time – one can always imagine what has actually happened – but allows other possibilities.

Again, the further along the axis one goes, the more recognizably the concepts one is dealing with are property concepts (see Section 6.2.2.3) and not object concepts (Section 6.2.2.1) or action/event concepts (Section 6.2.2.2)[27]. They point less toward things outside the conceptual space and more toward things, like themselves, within that space.

The vast majority of the property concepts are only implicitly higher order: e.g., concepts like **HEAV(IL)Y**, **DENSE(LY)**, **BROWN**, **QUICK(LY)** and so on. Those that have traditionally been classed as physical property concepts are, probably, more toward the first order, those traditionally classed as mental property concepts more toward the second order (see the discussion of property dualism in Section 4.2.1). A few concepts are explicitly higher-order. These include all the concepts of concepts: the concept of a first-order concept, the concept of an object or action concept, the concept of concept itself.

6.2.2 The *Other* Description: Sets of Logical Relations

Besides being well-behaved shapes, concepts are also, according to the unified conceptual space theory, sets of logical relations. What this amounts to is that, in addition to proximal connections between contiguous points of the space, and distal connections to referents (where the referents happen also to lie within the unified space: e.g., the concept of a first-order concept), concepts may map to distal parts of the unified space in a number of additional ways. Gärdenfors (2004) hints at this but does not go into any detail. I will begin to do so here.

All concepts, in addition to the proximal connections between contiguous points, and the distal connections to their referents, map to distal parts of the unified space in two additional ways:

Concepts possess integral dimensions. In this way, for example, **COLOUR** maps to **HUE**, **SATURATION**, and **BRIGHTNESS**. Integral dimensions are *necessary* (a colour *must* have a hue and a saturation and a brightness) but not ordered in any way (there is no priority among hue, saturation, and brightness.) I will henceforth refer to these as *parameters*. A concept inherits parameters from its superordinate concept(s) but must in some way either alter or add to those parameters so as to distinguish itself from the superordinate concept(s). Note that the parameters define a conceptual space of their own, whose dimensionality and boundedness depend on the number and nature of the parameters: e.g., for **COLOUR**, the parameters define a cone-shaped region of alternative possibilities. Indeed, this is the way the term "conceptual space" is defined by Gärdenfors (2004).

Concepts have associated with them various contextual elements: that is, things that the concept is co-present with in different contexts. If the contextual element is associated with the concept in a majority of contexts, then one can say it is typically associated with that concept. For example, cats typically meow; but my cat Kali does not, or cannot. Kali is still a cat. Kali does purr; but another cat, for whatever reason, may not. Therefore the **CAT** concept maps to the contextual elements **MEOW** and **PURR**. As the examples imply, contextual elements are not necessary; neither need they be ordered in any way. (One contextual element may be more likely than another, or they may be equally likely or unlikely.) Although the presence of any one or another contextual element is optional, a sufficient number of them may limit the possible scope of a particular concept to a specific instance of that concept. A concept cannot be given coherent interpretation except

[27] This fits in nicely with various evolutionary accounts of cognition, which see these sorts of concepts arising first: see e.g. (Donald, 1993; Torey, 2009) as well as Section 4.4.

with relation to some non-empty set of contextual elements, which collectively define its contexts of encounter and application. Henceforth I will refer to these simply as *contextuals.* Like parameters, contextuals inherit; however, unlike parameters, the inheritance for contextuals can be overridden: **BIRD** maps to the contextual element **FLY**, but **OSTRICH** does not. Because the contextual relation is customary rather than necessary, contextual inheritance is always *ceteris paribus*.

In Section 4.2.2, I distinguished three categories of concepts, according to their referents: object concepts, action/event concepts, and property concepts. Object concepts and action/event concepts, to the extent they are nearer the physical/concrete rather than the mental/abstract end of the axis of abstraction, will map to distal parts of the unified space in a third way:

Physical/concrete object concepts and action/event concepts, but not mental/abstract ones nor properties, will decompose into parts. Unlike parameters or contextuals, these will be ordered (see below); and one or more of them will be necessary. Henceforth I will call these *components.* Necessary components inherit: **MAN** has the necessary component **HEAD** because **MAMMAL** has the necessary component **HEAD**.

Details for all three categories follow.

6.2.2.1 Object Concepts

On the account I have been sketching, object concepts (see also Section 4.2.2.1) are, in general and in the first instance, first-order concepts; or, more accurately (since I am positing a continuum), they are *more toward* first-order than higher-order. At the same time, as I argued in Section 4.1.1, the conceptual agent capable of grasping higher-order concepts will find herself unable to make a principled distinction between first- and higher-order concepts: e.g., one's **CAT** concept becomes integrated with and inseparable from one's concept of one's **CAT** concept (in part because that is what it becomes as soon as one begins to examine – i.e., reflect upon – it). So, reflecting on an object concept *shifts it toward* being higher-order.

The components of object concepts are ordered in conceptual space analogously to how their referents are meant to be ordered in the referential space. That is, the components cannot be assembled in "just any order" for the same composite to result. *Furthermore, the components of object concepts will also be object concepts, of the same basic nature*: *like decomposes into like.*

Consider the concept of a man: the (minimum) necessary components relative to any particular application of the concept, and subject to revision, might be a head and a torso. A man with no arms or legs is still a man, but a man with no head or no torso is, at least in most instances, something else, albeit man-related: e.g., a corpse. Which precisely the necessary part(s) is/are is not important: those can at least in part be pragmatically determined by present context and subject to change. (If one meets a living man who *is* just a head and nothing more – attached, say, to a prosthetic robotic body – one might well revise one's concept to incorporate this new possibility.)

The parameters of object concepts will be property concepts. Just as properties describe objects, so property concepts describe (provide the integral dimensions for) object concepts (and, indeed, all concepts).

The contextuals of object concepts will split into two groups: *object concepts and action/event concepts.* On the one hand, an object shares a certain (physical or conceptual) space with other objects; on the other, an object has associated with it or is involved in certain actions (relative to an agent) or agent-less events.

6.2.2.2 Action/Event Concepts

> I soon discovered one reason why researchers may have neglected studying action concepts. They are difficult to define and they were, at that time, difficult to use as well controlled stimuli on a computer (Hemeren, 2008, p. 5).

Again, action concepts (or, as I prefer to refer to them, action/event concepts; see also Section 4.2.2.2) are, like object concepts, in general and in the first instance, first-order concepts. If they are more simply structured than object concepts it is, perhaps, because they map to entities in a temporal *plane* rather than entities in a physical *space* or *volume*. If concepts are, as I argued in Section 3.2.4, dynamic entities, then action/event concepts bring that dynamics front and centre – which may be partly what makes them difficult to define or study. At the same time, my intuition – which would be interesting to test empirically – is that action concepts are relatively more stable (resistant to change) than object concepts.

The components of action/event concepts are ordered in conceptual space analogously to how their referents are meant to be ordered in the temporal plane. Again, the components cannot be assembled in "just any order" for the same composite to result (in some instances, for *any* composite to result). The components of action/event concepts will also be action/event concepts, of the same basic nature.

Consider the concept of pitching as a type of throwing. The minimum necessary components of it are, perhaps, an aiming, a shaping of the hand, a drawing back of the arm, a snapping forward of the arm, and an opening up of the hand, in that order. All save the aiming are inherited from throwing: **PITCHING** minus **AIMING** roughly equals **THROWING**.

Action/event concepts will also have parameters, which again will be property concepts.

The contextuals of action/event concepts will split into two groups: *other action/event concepts on the one hand, object concepts on the other.* Actions and events do not take place in isolation but in the context of other actions and events. They take place as well in a context of the agents initiating the actions and the entities with which the action/events interact.

6.2.2.3 Property Concepts

If action concepts are, perhaps, somewhat more simply structured than object concepts, then property concepts are much more simply structured again. If object concepts and action/event concepts tend toward first-order, then property concepts tend toward higher-order: they are more abstract, more clearly mental (see Section 4.2.2.3). They are simplified from other concepts in two significant ways:

While other concepts are associated sets of well-behaved (normally convex) shapes across multiple domains, property concepts are generally specific to a single domain. This was discussed earlier (Section 6.1), where it was noted in passing that all concepts can be reduced to or treated as property concepts. That is to say, the nature of all concepts is to describe the nature of things (the things they are about), just as "heavy" describes the weight of something, or "brown" the colour of it.

Unlike other sorts of concepts, property concepts, at least in general, have no components. One would not normally, for example, have a concept consisting of being green and being heavy, or being young and being talented: that is, pulling different domains together (which is another way of making the previous point). Property concepts are defined solely by their parameters and contextuals.

As suggested earlier, to the extent that object concepts or action/event concepts are abstract, they will look increasingly like (or shade into) property concepts. What this means to the present discussion is that, at the same time, they will be increasingly unlikely to be describable in terms of components. Perhaps one can talk meaningfully about **MIND** having components such as e.g. **SELF-CONSCIOUSNESS**, **CONSCIOUSNESS**, and **SUBCONSCIOUSNESS**. But I think it makes more sense to talk of **MIND** as a property of brains or of agents, and **CONSCIOUSNESS**, etc., as a property of **MIND**. This brings me to the next point.

The parameters or integral dimensions of property concepts are themselves property concepts. Because concepts are not defined by their components, their parameters are, perhaps, heightened in importance.

The contextuals of property concepts are, on the one hand, other property concepts (where one finds e.g. **COLOUR***, one often finds* **MASS** *and* **DENSITY** *as well); on the other,* **either** *objects of which they are properties,* **or** *action/events of which they describe the properties.* No property attaches both to an object and to an action/event.

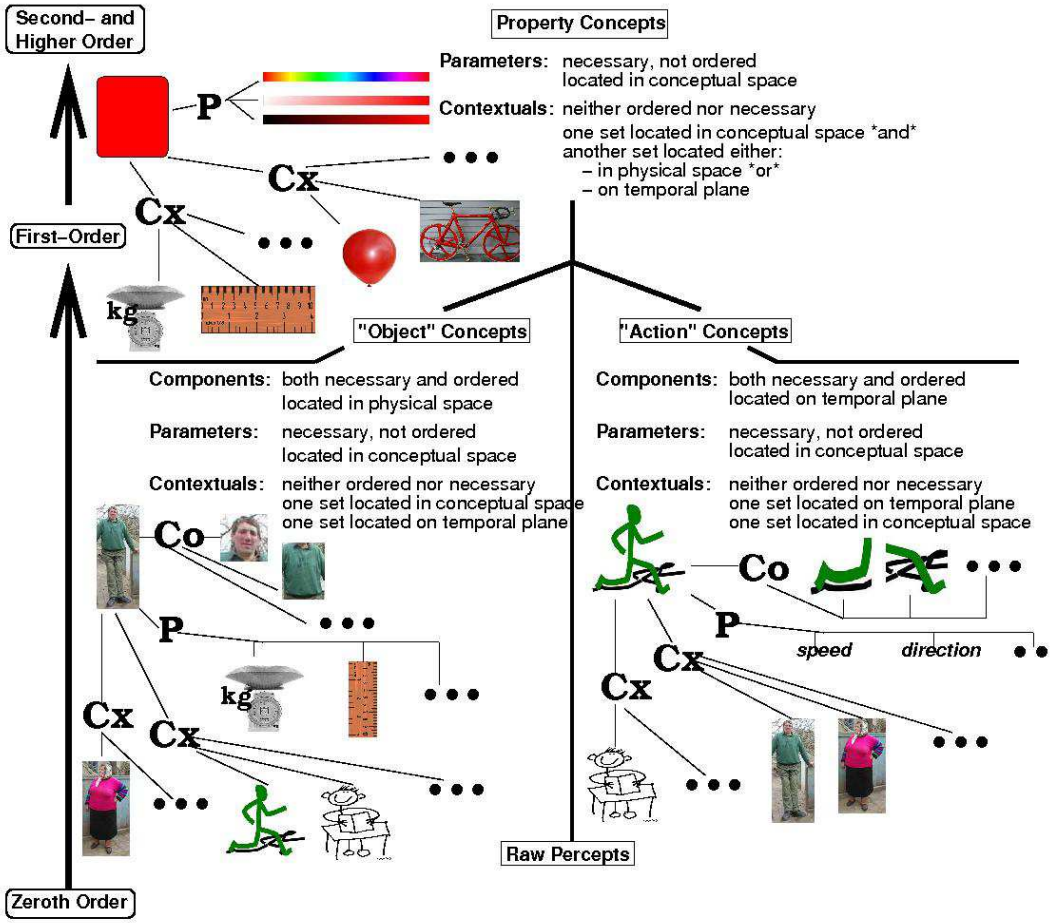

Figure 6.7: Varieties of concepts in the unified conceptual space, and how they are structured (**Co** = components, **P** = parameters, **Cx** = contextuals).

Before I continue, Figure 6.7 summarizes and offers a visual description of the main points of this section, showing how the three main categories of concepts are variously structured.

6.2.3 Limitations and Exclusions

The presentation of the proposed extensions and revisions to conceptual spaces theory is necessarily brief. It could turn into an entire thesis on its own. Some details have been omitted, and many more have yet to be worked out. This brings up the second point.

Although earlier versions of these ideas have been implemented in e.g. a writing environment for short story design (Parthemore, 1990) as well as experimented with in several smaller software projects, all of that work predates my exposure to Gärdenfors' research. The model as presented here has only been implemented in the relatively simple and incomplete application described in Chapter Eight. Putting theory into practice is, perhaps, the best way to discover what can and cannot work. Although the tension between theory and empirical study may always, ultimately, be unresolvable, making a tight loop of theory-model-implementation-theory is, as I argued in the first part of the chapter, the best strategy to address it.

I talked in Section 6.1.2, as I have elsewhere in the thesis, about the importance of embodiment. Both conceptual spaces theory and the unified conceptual space theory are, on their own, supremely

disembodied. They are, on their own, static as well. Although they refer to dynamic processes, they have no intrinsic dynamics. Though they are put forward by and debated by biological organisms, they are not directly *part* of any biological organism, continuously interacting with the environment in which it is embedded to create an ever-shifting conceptual space that is neither agent nor environment. The unified conceptual space theory does nothing, on its own, to change this deficiency – if such it is – in conceptual spaces theory. The concluding sections of Chapter Eight will offer some thoughts on how this might best be addressed.

(Neither does the unified conceptual space have any *internal* dynamics: the parts of the space are not in motion. This is a problem given that, on my account, concepts are in motion. As part of their formal treatment of conceptual spaces theory, Aisbett and Gibbon (2001, p. 217) attempt to describe such a dynamics.)

The questions of concept acquisition, application, and revision are deferred to Chapter Seven. But, although it will be touched upon there and some preliminary suggestions offered, the critical question of salience – that is, *why* some patterns are derived from regularities in the agent-environment interaction and not others; why some of *those* patterns give rise to concepts and not others; why, in other words, concepts get mapped to the referents that they do – will, for the most part, remain unresolved, with only the promissory note that the resolution lies outside the scope of this thesis.

Finally, the model of the unified conceptual space faces the twin dilemma of both not being algorithmic enough to permit direct translation into e.g. a computer model – the application in Chapter Eight is at best an imperfect translation; and, arguably, being already *too* algorithmic to capture some things one might reasonably want to say about concepts. Although I agree completely with Rick Grush and Pat Churchland (1995, p. 190) that it is unclear whether *anything* in the universe cannot (in principle) be described algorithmically, nonetheless some structures and processes will remain easier to describe algorithmically than others, and that will inevitably colour our perspective.

Still, remember my earlier point that a theory that tries to do everything does nothing well (Section 4.5). Remember also my warning that a theory of concepts pushed *too hard* toward completeness will become inconsistent in ways that are not the least innocent (Section 5.2.1.3). If concepts are themselves dynamic entities (Section 3.2.4), then a theory of concepts, as itself a conceptual entity, will be dynamic as well: a moving target, the best available answer to date (*relative always to some context of application!*).

Finally, if the perspective is biased one way or another, that may only be because, as I have argued, *all* perspectives are, by their nature, limited and biased. No perspective captures all perspectives, any more than any theory answers all possible questions. Instead, it generally chooses some hopefully useful and appropriate subset.

6.3 Conclusions

I have, in this chapter, set forth my best understanding of Gärdenfors' conceptual spaces theory. I consider it in the light not only of my own theory of concepts, as I am developing it in this work, but also in the broader context of competing theories of concepts, competing schools of philosophy, and empirical evaluation. In doing so I have strived – for all that conceptual spaces theory intuitively

appeals to me – to offer a critical evaluation, singling out those things I believe it gets right *vis a vis* e.g. the position of similarity in the order of explanation (always a tricky proposition for similarity-space-type theories), as well as those it arguably gets wrong: e.g., an over-reliance on representational language that may be more of a tactical than a theoretical mistake. Perhaps the theory's greatest contribution is in providing a neutral[28] language (the language of geometry) with which to break discussions about concepts out of the interminable debates over whether symbolic accounts or non-symbolic, association-based accounts of concepts or cognition are "right". The conceptual spaces theory is unabashedly pragmatic, denying that there can be any one single correct answer to questions like "what is cognition?" or "what is a concept?".

Gärdenfors' goal is not to tear down the opposition but to synthesize and ultimately co-opt it. He avoids introducing new terminology wherever possible and, despite his mathematical background, puts his theory in terms a non-mathematician and indeed, I think, any well-educated layperson can easily understand. It is an approach I would like to see more of in philosophy of mind.

In the second half of the chapter, I set out my proposed extensions to conceptual spaces theory in the form of the unified conceptual space theory, which describes how all of an agent's many conceptual spaces come together within a single space of spaces. I discuss this in the context of an individual conceptual agent but propose that a similar coming together of spaces must happen at the societal level, with no *a priori* precedence between individual and societal level intended.

In line with conceptual spaces theory, the unified conceptual space theory claims that concepts can be viewed as unstructured atoms (i.e., points), to which some arbitrary symbol or label may be attached; or as structured (non-atomic) entities. Such structure can be described in either of two complementary ways. The first is as shapes within the unified space, very similar to the well-behaved shapes in conceptual spaces theory (though with certain differences). The second is as bundles of parameters (integral dimensions); contextuals (elements of context); and, in the case of concepts of tending-toward-physical objects or actions/events, spatially or temporally ordered components. Object concepts and action/event concepts are more toward first-order, property concepts more toward second- and higher-order. There are, however, no fixed boundaries. Representationally, the first description is more iconic, the second more symbolic. Conceptually, the first is lower-order, the second higher-order. The first maps points in conceptual space to proximal points in that space, the second maps them to distal points.

Distal connections are, on closer inspection, of two different kinds. Those just mentioned capture what Frege called the *sense* of the concept: the manner in which it characterizes. But one may talk of what the concept is characterizing as well: its referent. At least part of the time, the referent of the concept will lie in the unified space as well. Of course, if one recalls the discussion of Section 3.2.1, like Fodor, though for very different motivations, I want to deny any substantive sense/reference distinction. What this means in practical terms relates to the argument of the previous chapter (Section 5.2.3.1) that the referents of concepts may never be reliably free of conceptual colouring, in which case, physical space (*as perceived*) and the temporal plane (*as perceived*) either overlap with or become themselves sub-regions of the unified conceptual space, and mind extends into world.

[28]...Neutral, that is to say, with respect to these debates.

Finally, I reflect on the limitations and shortcomings of the unified conceptual space theory in its present form. These include its lack of full implementation (e.g., in a computer or robot model) or any empirical testing, its seemingly disembodied and static nature (in contrast to my emphasis on embodiment and dynamics), its failure to address salience, and its necessarily limited scope of application – even while concluding that the last point is at least as much a strength as it is a weakness.

Though the next chapter will not resolve the issue of dynamics, it will go some way toward addressing it.

Chapter 7

The Co-Emergence of Concepts and Experience

The account of concepts I have offered so far has been ambitious, but it has had a striking omission: both the account of concepts and the concepts themselves have been given a largely static presentation. Chapter Two looked at the essential nature of concepts from a historical and contemporary point of view. Chapter Three considered their core properties; Chapter Five their limitations (as well as the limitations on any theories about them). Chapter Four attempted to place them into a context of agents and referents and – relevant to the present chapter – of application; but the discussion of application there was necessarily brief and preliminary, the attention elsewhere.

Chapter Six laid the groundwork for the present chapter, drawing together threads from the earlier chapters to show how Gärdenfors' conceptual spaces theory could account for concepts as both representations and (non-representational) abilities, and otherwise meet the basic requirements I had established for a theory of concepts; furthermore, how an extension of it I call the unified conceptual space theory could move it in an algorithmically amenable direction toward something more directly implementable in e.g. a computer model and therefore, one might reasonably hope, easier to test empirically. Indeed, one of the goals of this work is to put forward a theory that *can* be tested. At the same time, ironically, the unified conceptual space theory appears to do what Chapter Four said one should and indeed could not: to detach concepts from both the agents possessing and employing them and from any particular context of application, to turn them into some kind of free-floating entities and not things that are actively being lived.

If Chapter Six provided a bridge from the opening chapters to this one, then this chapter is a bridge from Chapter Six (the theory) to Chapter Eight (the intended application). Its goal is to tell a particular story of concept acquisition and application, as *conceptually distinct but logically unified* processes: two sides of one coin, two contrasting perspectives that we cannot quite conceptually unify. It will use the theory from the previous chapter for telling the story from both perspectives, one cognitively bottom up and the other top down. It will make much use of the twin concepts of *circular causality* (Section 7.1.2) and *co-emergence* (Section 7.1.3), both familiar from the enactivist literature, both involving the conceptual limitations and hiding the sort of paradoxes I discussed in Chapter Five. Finally, it will return us to the notion first suggested in Chapter Two: that representationalist and anti-representationalist perspectives are, on their own, critically inadequate; that whether concepts look more like representations or more like abilities will depend, finally, on where the observer stands.

7.1 Concepts and Experience: A Tangled Relationship

> We frequently compare the experiences we are currently having to memories of earlier episodes. Sometimes, we experience something entirely new, but most of the time what we see or hear is, more or less, the same as what we have already encountered. This cognitive capacity shows that we can judge, consciously or not, various relations among our experiences (Gärdenfors, 2004, p. 4).

I should begin by saying that combining bottom-up and top-down approaches to understanding cognition is far from a new idea. The approach described here shares something of the spirit, though not the style, of Antonio Chella *et al*'s (2000) approach to modeling dynamic scenes, which is more narrowly (and pragmatically) focused on visual understanding (much of it quite low level) and on the action portion of the conceptual space, as well as on how these could best be implemented in a robotic system[1].

Concepts and experience exist in a certain unavoidable tension. Experience is typically very much engaged in the moment and, in any case, grounded in the present. Concepts, on the other hand, abstract away from the immediate experience of the moment, stepping back from it to take a wider view. They have, as it were – to extend an idea from Lawrence Barsalou (Barsalou et al., 2007) who borrows it from William James – one hand in the past and the other in the future. Not only is it unclear whether "concepts" solely of application to the present moment would be of any use, it is unclear to what extent they would qualify as concepts at all.

At the same time, concepts and experience are critically dependent on one another. Unless one wants to postulate a large body of innate concepts – and even the latter-day Fodor (2008) seems reluctant to do that, at least in the usual sense of "innate" – most concepts will require experience to give rise to them. As I will argue in Section 7.2, that experience must ultimately be *sensorimotor-based* experience: the agent must be cognitively and physically engaged with its environment to experience either environment or self. The standard reference here, of course, is the classic study reported by Held and Hein (1963) in which kittens, deprived during their early development of the ability to interact actively with their environment, appear subsequently unable to make visual sense of that environment: they are effectively blind.

Experience is likewise dependent upon concepts. It is an open question and, perhaps, an unresolvable one, in what sense of the word a fully non-conceptual agent – one without any concepts or any degree of conceptual abilities – could be said to have experience. Perhaps a goldfish encounters the world in this way. In such an agent the "experienced" world would, in every instance, be entirely new. It might, in some limited fashion, have memory, as we think of it; but there would be no relating to the past *as* the past or to the future *as* the future, for that would imply some, at least minimal, conceptual abilities.

For the conceptual agent, such experience uncoloured by concepts (if experience it is) is, if I am right, no longer a possibility. Such an agent will never experience the "now" entirely on its own (though perhaps forms of meditation may get her closer than she would get otherwise, as may waking from a really deep sleep or from anesthesia, before the world resolves itself into its usual conceptual forms). Instead, the "now" is experienced in the light of past moments and in anticipation of future ones, as the opening quote from Gärdenfors suggests.

[1] The paper also provides an excellent literature review of work in the area of dynamic scene recognition.

Concepts reliably shape and re-shape our experience of the world: that is to say, they shape and re-shape our *world-as-experienced*. Though there is no reason to think that anyone is born with an innate concept of **DOORKNOB** – to borrow Fodor's (1998) example – once an agent has the concept then, in ordinary circumstances, that agent cannot fail to see a doorknob as a doorknob. Not only does the agent, in Fodor's language, become a reliable doorknob tracker; she cannot choose to unlearn it or even temporarily step aside from that role. In the language of someone like Alva Noë, once an agent has a sensorimotor profile of doorknobs, that profile is inextricably part of how that agent encounters and experiences her world, where "the sensorimotor profile of an object is the way its appearance changes as you move with respect to it (strictly speaking, it is the way the sensory stimulation varies as you move)" (Noë, 2004, p. 78).

Concept acquisition and application go hand in hand. Acquiring concepts is a process of applying concepts, which may themselves change in the process of acquiring new concepts. Within the context of the last chapter, our conceptual spaces, individually and collectively, are both the product of our interaction with our environment and the basis for it. To borrow a phrase from dynamical systems theory, they constitute dynamically coupled systems.

7.1.1 Dynamical Systems

> Perhaps the most distinctive feature of dynamical systems theory is that it provides a *geometric* form of understanding: behaviors are thought of in terms of locations, paths, and landscapes in the phase space of the system (van Gelder and Port, 1996, p. 14).

According to dynamical systems theory, "the cognitive system is not a computer, it is a dynamical system" (van Gelder and Port, 1996, p. 3), where time and change over time play key roles. A given system is a dynamical system if it is describable in terms of:

- A set of *state variables*, corresponding to a set of real numbers or points in *state space*, whose changing values determine the present state of the system.
- A set of parameters or *fixed points.*
- Either *differential* or *difference equations* for relating them, depending on whether time is being considered in a continuous or discrete manner.

Dynamically coupled systems are ones that interact in sufficiently rich ways as to constitute a single, joint dynamical system.

For all of the common ground – notably, the common preference toward describing cognition in terms from geometry – conceptual spaces theory and dynamical systems theory must part company at certain points, particularly the tendency by many dynamical systems theorists to think that *all* cognition can and should be described in terms of dynamical systems, and to deny any role to "symbolic" or "computational" accounts. Consider:

> . . . Dynamical and computational systems are fundamentally different *kinds* of systems, and hence the dynamical and computational approaches to cognition are fundamentally different in their deepest foundations (van Gelder and Port, 1996, p. 10).

Or:

> According to this ambitious doctrine the domain of the computational is empty, and dynamical processes will *eliminate* their computational competitors across all aspects of cognition (van Gelder and Port, 1996, p. 31).

More to the point, dynamical systems theory provides the wrong sort of tools with which to separate concept acquisition from application, and to tell a linearly ordered and causal story for each, as one might reasonably want to do. For that, we need a different tool, albeit one favoured by some of the same theorists: i.e., *circular causality.*

7.1.2 Circular Causality

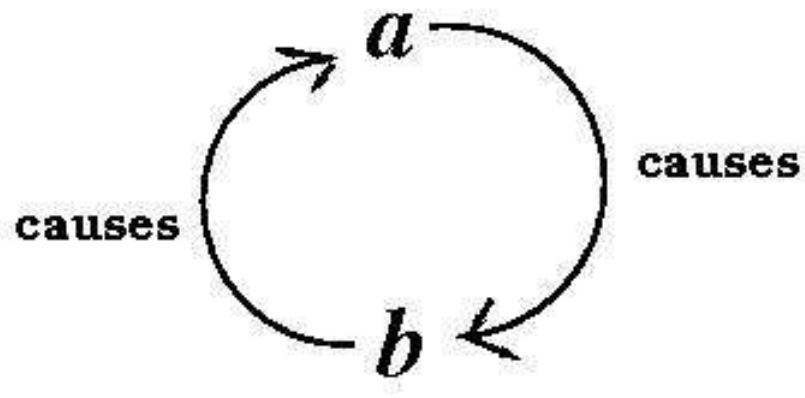

Figure 7.1: A model of circular causality.

Figure 7.1 presents the most basic model of circular causality: experience (a) gives rise to concepts (b), which in turn give rise to experience (a).

Although philosophers – particularly, perhaps, those within enactive philosophy[2] – often talk about circular causality as though it were just an alternative account of causal relations, it is not; at least, it is not that simple. Conventionally speaking, "cause" and "effect" are different things, related both by order (cause is supposed to precede effect) and necessity (the cause is *sufficient* for the effect *necessarily* to follow). In circular causality, what is cause and what is effect are purely matters of perspective, and every effect is, by tighter or wider loop, its own cause. If this sounds paradoxical, it should. Indeed, it should bring to mind the talk of eternally receding targets from Chapter Five.

I should not be understood as favouring a traditional, linearly causal model. Linear causality has its own issues, as David Hume (2003) is perhaps most famous for pointing out – most particularly for where the necessity relationship lies: between cause and effect, or only (as Hume himself concluded) in the perspective of the observer. It might seem that one is presented with nothing more than reliable co-occurrence (b has always been observed to follow a) and assigns it the status of logical implication (if a then b; a, therefore b).

[2] I have in mind someone like Thompson (2007), though the idea is important in (Varela et al., 1991) as well.

Neither am I rejecting circular causality, which I require for the account I wish to offer of concept acquisition and application. I am only pointing out what should, perhaps, be obvious: that the two perspectives offered by Figure 7.1, that a causes b on the one hand and that b causes a on the other, only make sense if considered separately from each other: *i.e., as two independent instances of linear causality*. Considered jointly, *together with how causality is being defined in the first place*, no consistent interpretation can be given.

Figure 7.2: A dragon swallows its own tail. (Photo downloaded from Wikimedia Commons: http://commons.wikimedia.org/ and edited.)

A circularly causal account may be very useful – even, in some cases, a necessary fiction; but push any such account too far, and one will derive a contradiction (see the discussion in Section 5.2.1.3, and compare Figure 7.1 with Figure 7.2). Consider the science fiction scenario of the person who travels back in time to become his own father (and therefore his own paternal grandfather, and his own great-grandfather, etc.). Although it may make for good story telling, it is impossible to make coherent sense of it.

7.1.3 Co-Emergence

So, we are left with the perspective toggling introduced in Section 5.4, where each perspective is like one half of the circle in Figure 7.1. I can account for the emergence of concepts from experience (Section 7.2), and I can account for the emergence of experience from concepts (Section 7.3). I can, and will, argue (Section 7.4) that no priority can logically be made between the two accounts (even though a linear text insists that I put one before the other, and both logic and convention suggest placing acquisition first: one cannot make use of what one does not possess). What I cannot do is offer one unified account, based on one unified perspective. That is not, however, the fault of the account.

Concepts and experience are *co-emergent*: each (to borrow Varela's phrase) "brings forth" the other. It's like the chicken-and-the-egg problem: which comes first? Something, of course, must

logically start things off, but it need not be either concepts or experience as we understand them; especially if, as I have been suggesting, each presupposes the other. Caught within our conceptual perspective (see Section 5.2.1.2), we cannot step outside of it: we cannot simply put our concepts or our conceptually structured experience aside.

7.2 Concepts Emergent: The Acquisition Story

> What transforms the "booming, buzzing confusion" that enters our eyes and ears at birth into that orderly world we ultimately experience and interact with (Harnad, 1990c, p. ix)?

Concepts – as described by Fodor or Prinz or Millikan or Gärdenfors or pretty much any theorist of concepts – are paradigmatically abstract and cognitively high level. If this is most clearly true on the concepts-as-representations accounts, it is only marginally, if at all, less true on the concepts-as-abilities accounts.

Sensorimotor engagements are paradigmatically concrete and cognitively low level. On the face of it, the two could not be further different and removed from each other.

Yet it would be a mistake, I believe, to take these either as fully independent systems (such that an exploration of the former could take place without any consideration of the latter)[3]; or as semi-independent and interacting systems needing somehow to be fitted together. At the same time, the one cannot be eliminatively reduced to the other, as some dynamical systems theorists are inclined to do (see Section 7.1.1), and as Vittorio Gallese and George Lakoff favour in (2005). Rather, they are positions toward either end of a continuum, neither of which can be divorced from the other; and experience straddles both.

Mental content and meaning do not "come for free": at some point they must begin with, or be caused by, that which is not mental content, that which does not already have meaning. That is to say, they must be *grounded*. Noë (2004), Harnad (2007), and others have written that all mental content, conceptual or otherwise, must be grounded in specific sensorimotor engagements, and so sensorimotor engagements are partly constitutive of that content. I agree. Note that this makes mental content not fixed but contingent: *if* I do this, *then* I will experience that; *if* the agent does this, *then* the agent's mental content will be that. Mental content is defined out of interaction; there is no logical separating of input from output, as the dynamical systems theorists would strongly concur. This is why one talks, not of a sensory system and a motor system, but of a *sensorimotor* system. Note, too, that this is consistent with (indeed should be seen as a refinement of) the classical empiricist tradition that grounds cognition in experience.

[3] I have in mind someone like Fodor.

7.2.1 Noë's Sensorimotor Account

> . . . We ought to reject the idea – widespread in both philosophy and science – that perception is a process *in the brain* whereby the perceptual system constructs an *internal representation* of the world (Noë, 2004, p. 2).

Much of what follows is consistent with Noë's brand of enactivism, as noted in Section 6.1.2. Nonetheless, I do take issue with Noë's account at a number of points[4].

- *Noë favours a more linearly structured account than I.*
  Noë's attention in his 2004 book is on how cognition consists of or is built upon continuous sensorimotor engagement. Perception is never truly passive but should be understood, even in its apparently passive instances, as an engaged physical activity. This is important, and either downplayed or missed entirely in many older accounts of cognition, which are not unfairly caricatured as over-intellectualizing matters. But one could equally turn the perspective around to look at how sensorimotor engagement consists of or is built upon perceptions: how, as a matter of principle (on anti-realist accounts) or practice (on certain realist accounts), the mind constrains the world. To wit: Noë's causality is linear, from sensorimotor engagements to sensorimotor profiles to higher cognition; my preferred model of causality, as noted above, is circular.

- *Noë's account is overly forward looking.*
  I believe that Noë's account is much more forward- than backward-looking: Noë has a lot to say about where expectations take us but relatively little to say about where they come from (as opposed to e.g. Chrisley's Expectation-Based Architecture approach (Chrisley and Parthemore, 2007b)). This might in part be a consequence of Noë's strong externalism; talking about where expectations come from might seem to require a more internalist-looking perspective, one that has more to say about the mental life of the agent.

- *Noë's account is strongly externalist; I wish to reject the extremes of both internalism and externalism.*
  As Anthony Morse and Tom Ziemke have written (2010), Noë focuses on sensorimotor contingencies to the exclusion of any consideration of the agent's bodily states in general or emotions in particular. Without an account of emotions and, consequent to that, of motivations, Noë has no explanation for why some affordances in the environment are salient and others not. This follows directly from Noë's externalism. The sort of enactivism I have endorsed (see Section 6.1.2) views either internalist or externalist perspectives, when taken on their own, as deeply misleading. One of the important things to remember about affordances is that they are not in the environment (as Gibson (1986) is often read) but (metaphorically speaking) in the interaction, always relative to the perspective of the agent at a particular moment in time (as a more nuanced reading of Gibson might allow).

- *Noë's account is strongly anti-representationalist, whereas I favour a qualified representationalism.*
  Finally, and perhaps most critically for my account, is the reason why I classified Noë (see Section 2.5.3) on the "abilities" side of the abilities vs. representations divide: Noë eschews representational language. As should be clear by now, I take that to be a mistake. At the same time, I want to endorse fully the quote of Noë's that began this section.

[4] Much of the content of this section I owe to informal discussions with Anthony Morse, my co-author on (Parthemore and Morse, 2010).

7.2.2 More General Issues With Sensorimotor Accounts

There are further problems, not particular to Noë's account, but common to any accounts that might be seen to focus too narrowly on sensorimotor explanations. The difficulty is how one gets beyond specific sensorimotor engagements: how one generalizes to the sensorimotor profiles needed to explain specific affordances, nevermind abstract conceptual thought. This is what Harnad has called "the problem of extracting reliable categories from experience" (1990a, p. 538).

Take for example the account of Gallese and Lakoff (2005), according to which the most abstract of concepts is, on any occasion one can name, no more than a specific sensorimotor engagement, albeit with parts of it (e.g., the full activation of the motor cortex, with consequent movement) suppressed. The concept they focus on is that of grasping. The similarity between their account of grasping and George Berkeley's account of triangles (1999) is striking. For Berkeley, the general concept of a triangle is nothing more than a specific instance of a triangle with details suppressed: the length of the sides or the measure of the angles. For Gallese and Lakoff, the general concept of grasping is likewise a specific instance of grasping with details suppressed: i.e., the "actual" carrying out of the action.

One might be tempted to cede the point for grasping and reserve concerns for concepts like that of democracy or enlightenment, or such extreme abstractions as the concept of a concept itself; but that would, I think, be a mistake. To the extent that one's concept of grasping is a representation of grasping – and I have argued that, for the conceptually reflective agent, all concepts take on this representational aspect – the representation may have as much (or as little) to do with the represented as e.g. a representation of a dog (such as a painting) has to do with a dog. I have further argued that the role of all concepts is to simplify (Section 5.2.2) and to abstract away from the particulars of context, from (*pace* Gallese and Lakoff) any particular application; in which case, some additional mechanism or mechanisms are needed. I propose that the unified conceptual space meets this requirement.

Of course I, too, want to reject a sharp class/instance distinction, but not by merely eliminating it. By trading on an essential ambiguity – according to which all instances can (within practical limits) be treated as classes and all classes can be treated as instances of more general classes (see Section 6.1) – I will show how one can get from either one to the other, on whatever cognitive level.

7.2.3 An Alternative Account: Sensorimotor ++

> How can one *generalize* from single observations to general laws (Gärdenfors, 2004, p. 205)?
>
> The ANN [artificial neural network] model I outline here is based on Kohonen's... *self-organizing maps.* The purpose of the method is to *reduce the representational complexity* of the input in an efficient and systematic way. Thus the proposed method can be seen as a way of answering [the] question... on the subconceptual level (Gärdenfors, 2004, p. 221).

As the unified conceptual space theory was presented as an extension of Gärdenfors' work, so the position I would like to take on sensorimotor grounding of cognition should be understood as an extension of Noë's work, whilst addressing the above reservations. One might be tempted to call

it "sensorimotor plus plus"[5]: sensorimotor engagements ***plus*** somatic and other bodily information (*per* Damasio (2000) and Morse and Ziemke (2010)) ***plus*** (with appropriate qualifications) representational language, as located within a conceptual spaces framework.

As Gärdenfors' diagnoses, the emergence of cognition in general and concepts in particular from their pre-conceptual and pre-cognitive origins is largely a matter of induction. Although we may be most comfortable thinking about induction as a very abstract and high-level cognitive process, an analogous process must logically, it seems, be going on at the lowest levels of cognition. Gärdenfors couches his account of low-level induction in terms of Kohonen maps, a special kind of artificial neural network that preserves the topological relations of the input space. The details of Kohonen maps is not important here. I want to offer a different but, I hope, complementary account, first abstractly and then through the specific mechanism of the unified conceptual space theory.

As on standard connectionist or associationist accounts, the story begins with pattern recognition, with the caveat that there should be some, minimal, pre-existing ideas in the system about what sort of things patterns are (in the most general possible way). These constitute a minimal set of primitive building blocks – proto-concepts: mental structures that are concept-like yet fail to meet one or more of the intrinsic properties of concepts, such as being under the agent's endogenous control. Such structures should be governed by nomic relations (of the kind endorsed by Fodor (1998) for concepts in general) rather than experientially derived. In terms of conceptual spaces theory, they provide an initial partitioning of the unified conceptual space (see Section 7.2.4), one from which all subsequent partitioning is derived, through a process of recognizing regularities in the perceptual stream between one moment and another and another: repetitions that are recognized as somehow salient.

I will not attempt to offer any proper account of salience. Such an account lies far beyond the scope of the present work. Nonetheless I find intuitively appealing the attempt by many in the enactivist camp – recently e.g. by Evan Thompson and Mog Stapleton (2009) – to ground minimal salience in the survival of the organism. What is salient is what enables the organism to survive, and other saliences should follow directly or indirectly from there. More of the detail can be fleshed out on a somatosensory account such as offered by (Morse and Ziemke, 2010).

The most basic perceptual regularity, then, is a mapping between a set of points in "raw" (but still proto-conceptually structured, and hence minimally interpreted) experience: something was observed, and then it was observed again, and again. To borrow an example from robotic vision, it could be something as simple as a recurring pixel at the same location in a visual display, or a sudden change of pixel value.

An agent can derive many such regularities. Just as one's "raw" perceptions define a perceptual space, so, too, these regularities collectively define a space of their own: a space of minimal regularities, albeit one that preserves something of the topology of the original perceptual space.

The trick is that, just as patterns could be discovered in the original perceptual space, so too, patterns can be found in the higher-level regularity space. The latter are patterns *of patterns*, regularities in the regularities.

[5]The name was suggested by Gärdenfors.

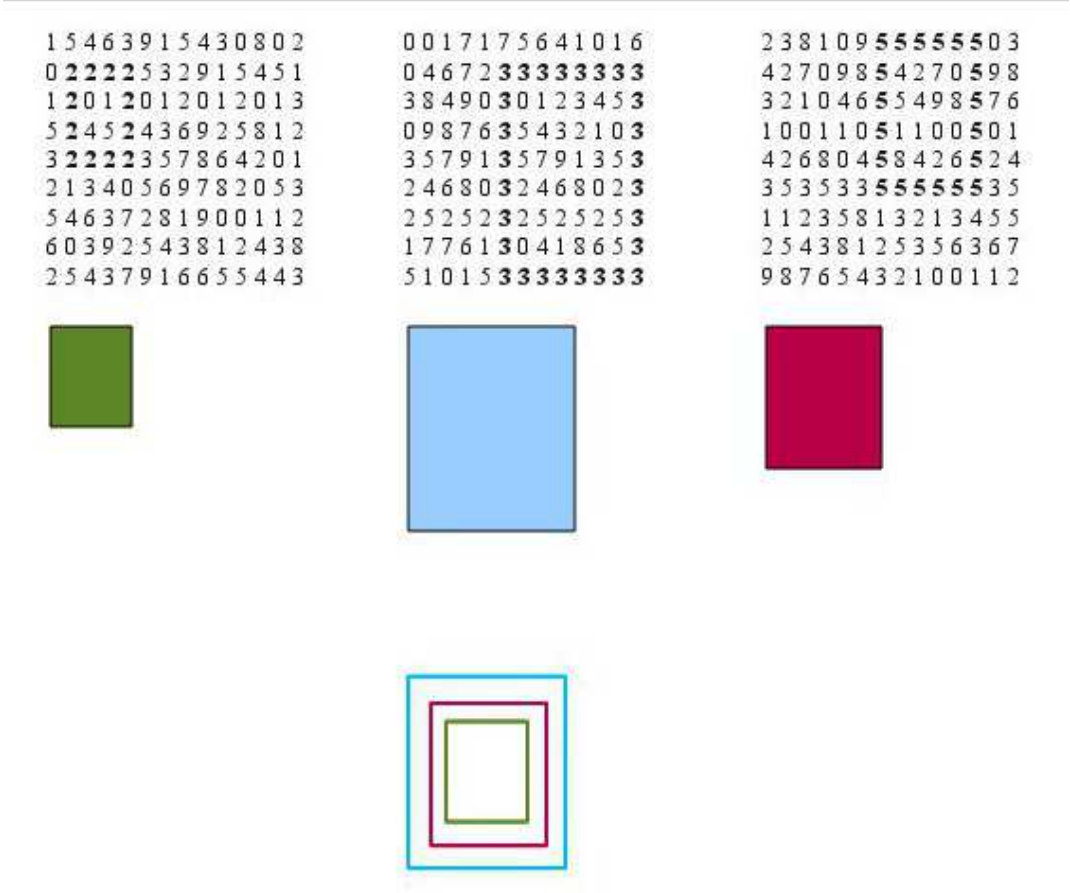

Figure 7.3: Patterns, patterns of patterns, and patterns of patterns of patterns.

Consider Figure 7.3. In the first row, digits are highlighted where there is a repetition of three or more of the same digit in any column or row of a grid of digits. In the second row, a second-order pattern, that of a common rectangular shape, is extracted. In the third row, the shapes are placed inside one another relative to their size: a *third*-order pattern. Of course, this same idea could be extended through any number of steps with other examples.

So regularities are found among the first-order regularities, yielding second-order regularities; regularities are found among the second-order regularities, yielding third-order regularities; and so on. At each level, the agent steps further back from the moment, and the "moment" itself (i.e., the minimal individuable unit of time) becomes more and more stretched out.

Regularities in the regularities, and regularities in *those* regularities, yield increasingly complex, increasingly abstract mental content that, as the process iterates over and over, eventually becomes recognizable as first- then higher-order concepts, iconic then symbolic representation, unconscious (non-introspectible) then self-conscious (introspectible) content. The result is an associational hierarchy that at its base is strongly associational and at best *very* weakly symbolic, at its summit strongly symbolic and at best *very* weakly associational.

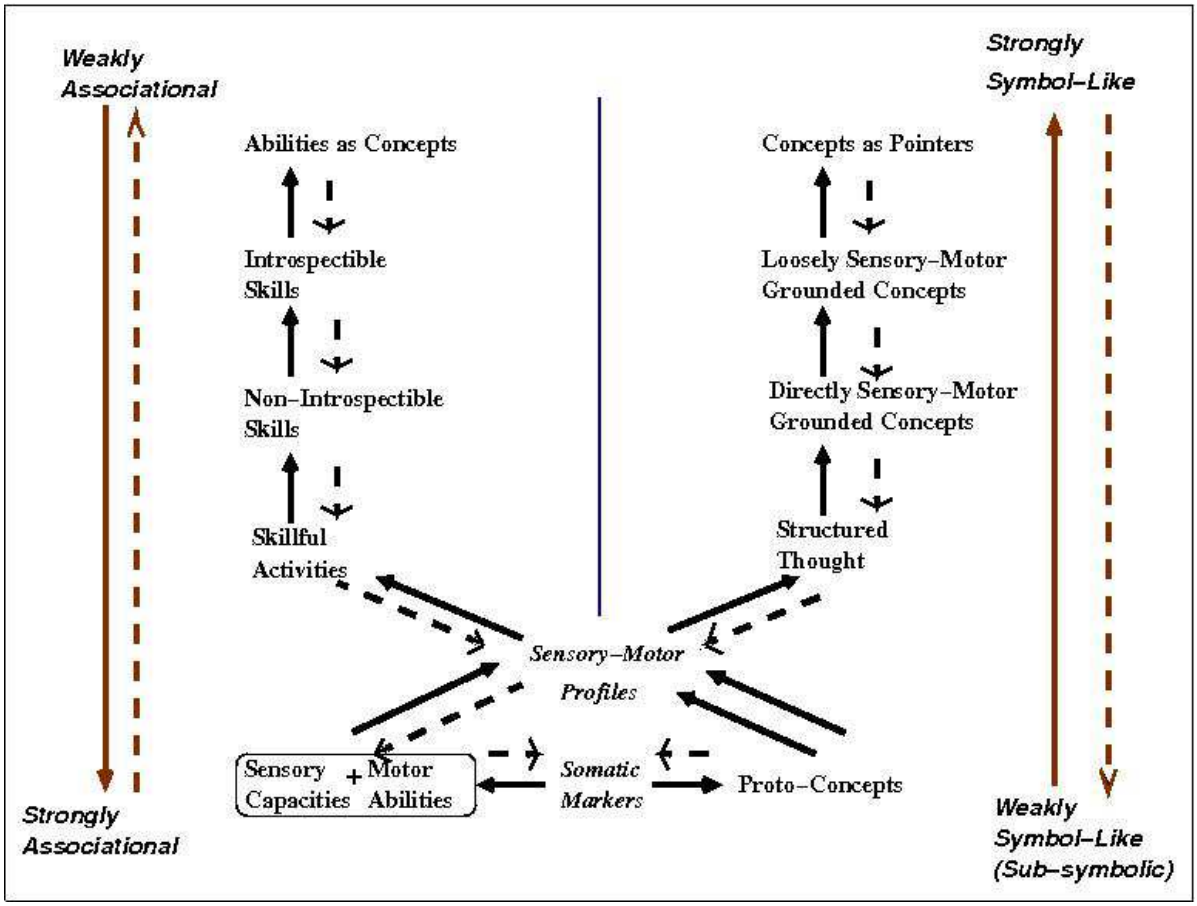

Figure 7.4: Perspectives on a cognitive continuum.

More accurately, the "hierarchy" should be understood as a continuum. At one end, one is, as it were, drowning in a sea of detail; go too far in the other direction, however – toward the very rarefied, the very abstract – and no useful detail remains.

See Figure 7.4: somatic markers (the term originates with Damasio (Damasio et al., 1991)) combine with, on the one hand, sensory capacities and motor abilities; on the other hand, proto-concepts. Together these give rise to sensorimotor profiles. From there, one can follow either of two paths, representing two contrasting perspectives. The narrow blue line down the middle represents the unresolvability of these two perspectives.

As one ascends through the associational hierarchy (following the solid black lines), the richness (dimensionality) of the referring structures is reduced at each step and the richness (dimensionality) of the referent structures (the target descriptive space) is likewise increased; compare Gärdenfors in describing the working of this process in the case of Kohonen maps:

> . . . Points closely related in the high-dimensional space are mapped onto closely related points in the low-dimensional space. Since dimensionality is reduced, this entails that regions of the high-dimensional [perceptual] space are mapped onto points in the low-dimensional [conceptual] space (2004, p. 222).

Look back at Figure 7.3: the second row simplifies and abstracts away from the first, the third row from the second. At the same time, the semantic content – the represented *pattern* – is enriched at each step.

In the limit, the referring structures come increasingly to look like arbitrary symbols, whose form bears no obvious relation back to any particular context; while the referent structures are maximally structured. The referring structures become mere pointers to richly structured sub-regions within the unified conceptual space. In this manner, the unified space is gradually transformed from a largely unstructured space to a space of many spaces.

If one inverts the associational hierarchy (following the dotted black lines downward), then what at first looked very much like symbols will shift into iconic representations and gradually lose themselves in context as their meaning becomes more and more defined by that context, until most (if not all) of what we understand by symbols disappears, and we are left with "bare" interactions.

I can now, finally, return to and refine the definition of "concept" first offered in Chapter One:

> A concept is a synchronized pattern of relatively abstract, relatively higher-order association between some aspect of the mental world of the agent ("self") and some matching aspect of her experienced environment ("non-self").

A concept is recognizably a concept to the extent that it abstracts away from the particulars of context, even as it is always then applied back to particular contexts. It both abstracts away from and is structurally isomorphic to its referent in perception. It need not necessarily correspond in any way to whatever caused that perception, only vary as whatever caused that perception varies[6].

I will return to and revise this definition once more before I close the chapter.

7.2.4 Partitioning the Conceptual Space

> Clearly, no organism is born a blank slate. Some categories are innate (Harnad, 1990b, p. 2).
>
> Presumably, the child masters only a few prototypes in animal space and these prototypes are used to generate a partitioning of the entire space. Consequently, the child will overgeneralize a concept in comparison to its standard use.... When the child learns more prototypes for other animal concepts, however, it will gradually adjust an early concept to its normal use since its partitioning of the animal space will become finer (Gärdenfors, 2004, p. 125).

The coming to be of a conceptual mind is the progressive partitioning[7] of a conceptual space that is, at the same time, a space of spaces: what I have termed the unified conceptual space. I have already suggested, at several points, that the conceptual mind does not start out as a *tabula rasa* but is already – albeit minimally – structured. This is by virtue of a small set of innate proto-concepts and a small set of "rules" (identifiable as such only by some theorist) for breaking them apart (to make more subtle distinctions), combining them together, adjusting their boundaries, and otherwise matching them against experience. Such is the symbolic (or rather, symbol-like) heart of what, at the most basic level of cognition, is otherwise an all but entirely non-symbolic

[6] Compare Gärdenfors (2004, p. 109): "... Representations need not be similar *to the objects* they represent. What is important is that the representations preserve the similarity relations between the objects they represent...."

[7] Compare Dominic Massaro's notion of *categorical partitioning* (1990). Note that, for Massaro, this partitioning is *not* already present in perception but is added later.

matter (see again Figure 7.4). The suggestions for proto-concepts are taken directly from the earlier discussions in sections 4.2.2 and 6.2.2. These are the great-great-great-etc. grandparents of all the other, "true", concepts.

It is important to emphasize, again, that proto-concepts are *not* concepts: that is, the structuring of the unified space begins at a level far below the conceptual. At the same time, proto-concepts are not entirely non-conceptual, either, nor is the conceptual space they structure: otherwise it would not be a conceptual space. As noted earlier, one need not accept McDowell's position that experience is *fully* conceptual "all the way out" to allow that it is conceptually *coloured* all the way out (see Section 5.2.3.2).

7.2.4.1 Initial Partitioning

This initial situation is described in Figure 7.5 (*cf.* Figure 6.6). Object- and action-type things dominate the space, all of which is compressed below the level of first-order concepts (but above zeroth-order).

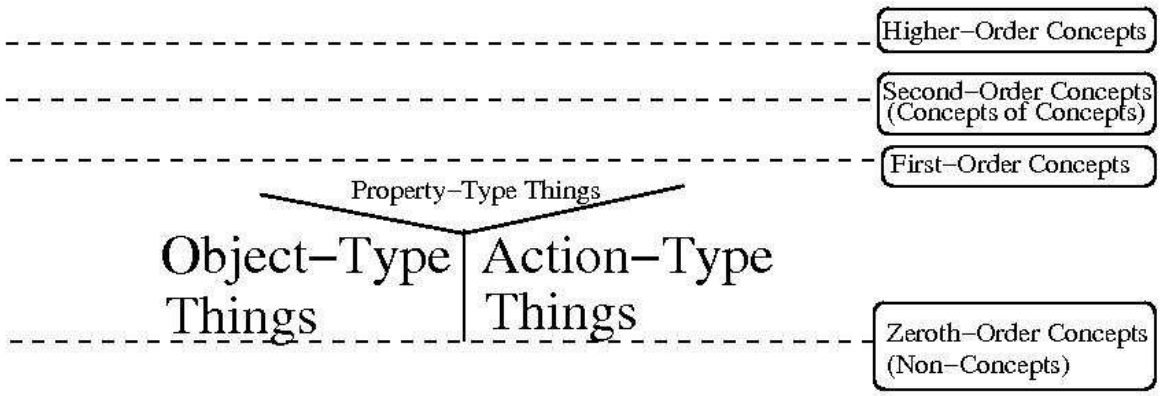

Figure 7.5: Initial partitioning: minimal structure.

A few caveats are in order. First, the distinction between zeroth-, first-, second-, *n*th-, and higher-order concepts should not, *per* Section 4.1.1, be taken as a series of discrete levels but rather rough positions along a continuum. Those positions relate to the degree of *explicit* reflection required to grasp the concept: that is, higher-order concepts, like the concept of a second-order concept, require not just thoughts about thoughts but some level of active awareness *of* thoughts about thoughts. Second, one might be tempted to substitute in Rosch's (1975; 1999) labels of *subordinate, basic-level,* and *superordinate* for *first-order*, *second-order,* and *higher-order*; but that would be a mistake, for although Rosch's labels do, indirectly, relate to how abstract categories are on a scale of concrete/physical to abstract/mental, they are first and foremost about the relationship between "parent" and "child" categories. That is, following the discussion in Chapter Six, they lie along the axis of generalization (Section 6.2.1.1) not the axis of abstraction (Section 6.2.1.4), allowing that those axes do, at one extreme, converge.

7.2.4.2 Subsequent Development

As the partitioning proceeds and the space expands into the area above first-order concepts, object-type things (now recognizably objects) and action/event-type things (now recognizably action/events) continue to dominate. See Figure 7.6. Particularly at the proto-conceptual level, the space becomes recognizable as a Voronoi tessellation (see Figure 6.1, which the proto-conceptual layer in this figure is meant to resemble). Initial tendencies will be to overgeneralize, as Gärdenfors reminds us (see the opening quote for Section 7.2.4). Notice that, in line with the comments about Rosch above, although conceptual development *does* proceed from the sub-conceptual to the conceptual to the meta-conceptual, it does *not* proceed from the subordinate to the basic-level to the superordinate. Rather, it jumps *first* to the basic-level and expands to subordinate and superordinate from there, as Rosch notes in describing some of her empirical results:

> ...Basic objects were shown to be the first categorizations made by young children, and basic object names the level of abstraction at which objects are first named by children and usually named by adults (1975, p. 587).

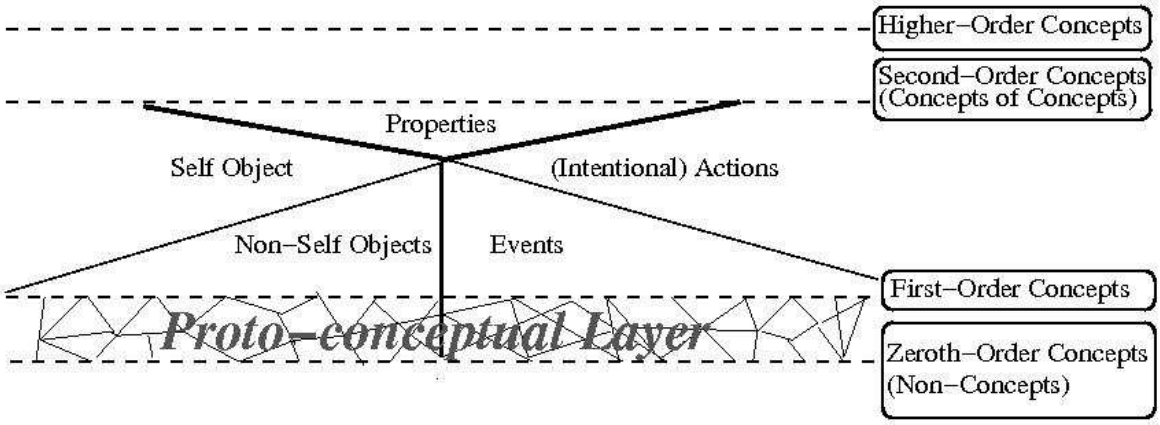

Figure 7.6: Initial first-order concepts and distinctions.

That said, one of the most basic "object" distinctions to be made is between self and non-self (see Section 4.1.3), and one of the most basic "action/event" distinctions is precisely between (intentional) actions and (non-intentional) events (see Section 4.2.2.2). These would seem, intuitively, to be foundational to all our other conceptual distinctions: so e.g., a child must be able to distinguish self from non-self before proceeding to recognize, or name, basic objects.

7.2.4.3 Advanced Partitioning

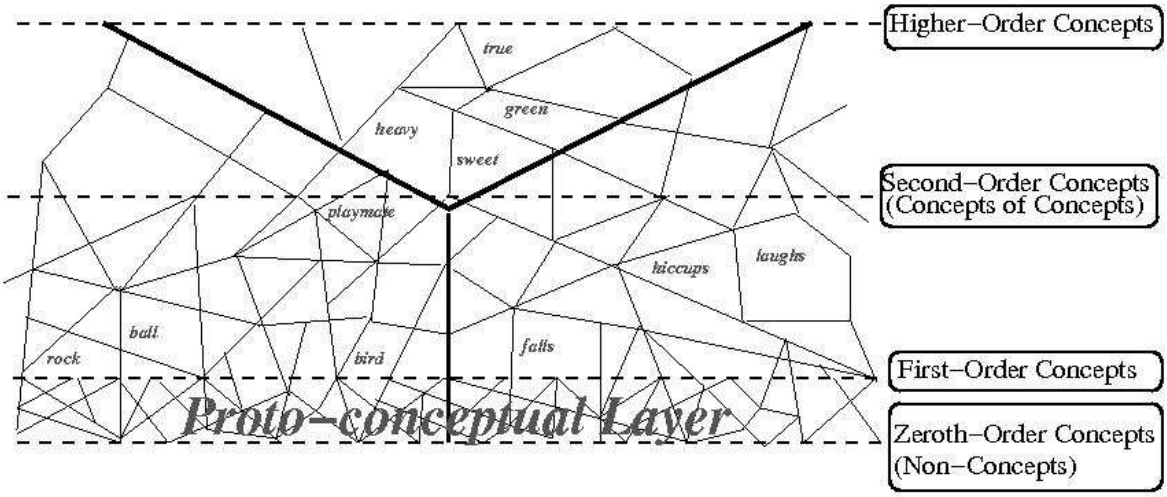

Figure 7.7: Higher-order concepts and further distinctions.

As meta-cognitive abilities develop (the capacity for thoughts about thoughts, and for thoughts about *those* thoughts), the partitioning expands further toward higher-order concepts, giving more and more of the space the appearance of a Voronoi tessellation (see Figure 7.7). Properties take on increasing importance, even while sub-partitioning of object and action/event subspaces continues. To the extent that the space is still largely unstructured, however, there will remain a consequent tendency to over-generalize. *Per* Gärdenfors' account, this problem will persist (and perhaps never entirely go away: note e.g. the categorization differences between bird watchers and bird "novices", or between expert mushroom gatherers and the greater number of people who cannot tell a mushroom from a toadstool).

7.2.5 Mapping Conceptual Space Onto Conceptual Space

> ...Humans have powerful abilities to detect multiple correlations among different domains.... In the theory of conceptual spaces, this kind of inductive process corresponds to determining *mappings* between the different domains of a space (Gärdenfors, 2004, p. 228).

The process I have just described, where the unified space begins as largely if not entirely unpartitioned and ends up intricately partitioned, is a movement from the maximally general (*something* is salient) to the maximally specific (the salient thing is e.g. my weight on the scales in my bathroom at 7:03 this morning). Here are concepts under what I termed the "first description" (see Section 6.2.1), as well-behaved shapes in the unified space .

However, there is a competing perspective that is just as valid, whereby concept formation is a movement from the maximally specific (applicable to the narrowest possible range of contexts) to the maximally general (applicable to the widest possible range). This is because, at the same time that the unified space is being partitioned, it is becoming more and more structured in another way, as parts of it map onto each other, both through conceptual reference, in those cases where the referent also lies within the unified space, and through filling in the details of components,

parameters, and contextuals. Here are concepts under the "second description" (Section 6.2.2), as "sets of logical relations".

These mappings are initially very limited, rendering the proto-concepts and their immediate derivatives applicable only within quite narrow contexts, things to be used and quickly discarded (or recycled). Over time they become increasingly sophisticated, allowing agents to detect "multiple correlations among different domains". So on one end of the continuum, one has maximally general proto-concepts that are applicable only within extremely specific contexts (everything is just a [unique] *thing*); at the other end, one has maximally specific concepts (relating only to one particular thing) able to locate their referents in the broadest possible range of contexts. Compare this to the earlier discussion in Section 7.2.3.

The description of space mapping onto space in this way, given in Section 6.2.2, was very high level. At its most basic level, it should begin with distinguishing proto-conceptual instances one from another by some proto-conceptual "understanding" of their different properties[8], using those properties to begin to organize them into a hierarchy of classes. Some of them will have parts (components) relating to each other in a predictable fashion. Proto-conceptual object-type things will get associated both with other object-type things and with various action/event-type things; the same will happen in reverse for proto-conceptual action/event-type things.

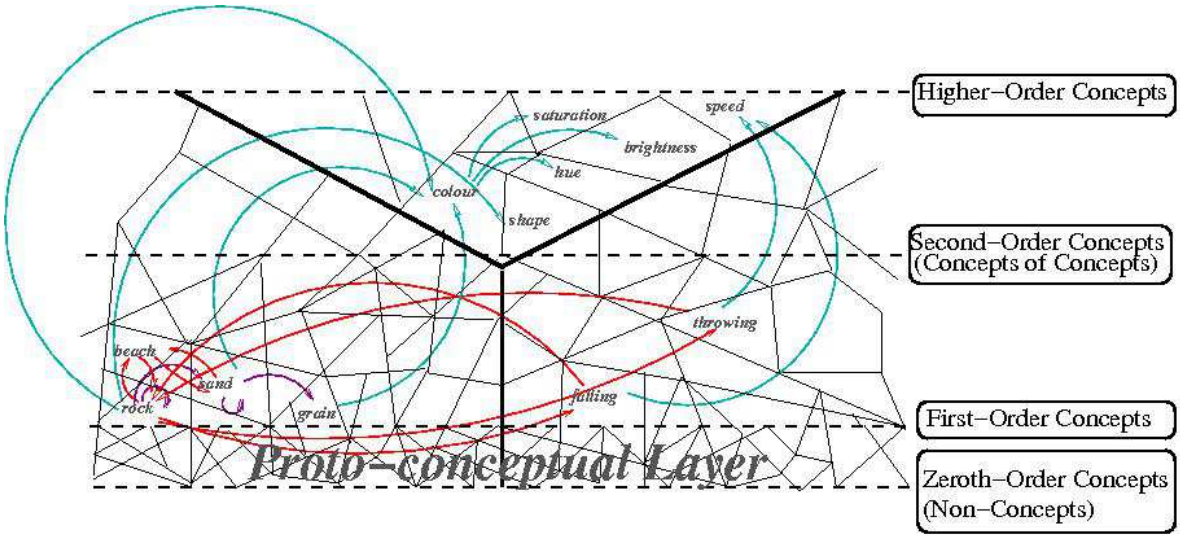

Figure 7.8: Mapping between different parts of the unified space.

Consider Figure 7.8. The purple lines represent components: a rock can "consist of" rocks or sand; sand consists of grains. Cyan lines represent parameters: both rocks and sand have colour, which itself has the parameters of hue, saturation, and brightness. Both throwing and falling have the parameter of speed. Red lines are associational: both rocks and sand are associated with beaches. Rocks, but not sand, are associated with falling and throwing, both of which are associated with (among other things) rocks.

[8]... That is, their parameters (or, in Gärdenfors' terms, integral dimensions).

7.3 Experience Emergent: The Application Story

> A brain is interposed between the stimulus and the response, totalizing the sensory input of the organism into an internal representation of what goes on in and about it (Torey, 2009, p. 6).

Of course if the account given so far – of concept acquisition – was all there was to be said, then concepts really would be static entities, with no means for update or obsolescence. But concepts have no value, no meaning, unless at the same time they are being acquired they are being applied. On the account I am offering, they are at least as much skillful abilities as they are expressible knowledge (*cf.* (Morse and Ziemke, 2007)). Furthermore, on any conceptual-spaces-derived account, they will be fundamentally *dynamic* entities.

To borrow a page from the classical definitionists (Section 2.3.1), the concept acquisition account just outlined can be turned on its head, verifying instead of discovering, disassembling instead of assembling, in the same spirit in which definitions are neutral as to whether they are defining new concepts or identifying and verifying old ones. Before, concepts were being abstracted away from experience, away from the particulars of the moment. Here, concepts are being applied back to experience, back to the particulars of the moment.

Unfortunately we do not have, as we had with Noë and the mostly bottom-up-driven process of concept acquisition, a similar guide to the mostly top-down-driven process of concept application. I will need, instead, to help myself to several diverse guides.

As do many in the enactive camp, I reject as not very useful (as a general model for cognition, even viewed top down) the cognitivist input-output-based model of cognition, exemplified by SMPA: *sense-model-plan-act*. Although there are, of course, many differences, what cognitivist approaches generally have in common is a splitting apart of sensory input from motor output and a treatment of higher-level cognition as independently explainable from lower-level details of "mere" implementation. That said, they remain surprisingly popular: witness the recent publication of Torey's book (2009), which ironically is, in many ways, much more traditional than it is revolutionary.

Torey's approach, as neatly summarized in the opening quote, is precisely the sort of representationalist account I do *not* want to give. Remember that representations as I intend them (see Section 2.6.4) are neither internal nor external, nor are "mental" representations mental in a way that other representations are not. Rather, representations are an ineliminable part of our first-person perspective. They stand not between the pre-experiential agent and his pre-experiential world but between the (self- or other-) *perceived* agent and the *perceived* world.

Questions about the nature and existence of representations have formed a running theme throughout this work. Representations are less ontological reals than perspectives we take – perspectives that, at the same time, we cannot simply set aside. Perhaps one should, in the end, talk not of representations but of a *representational stance.* I will return, one final time, to the question of representations in Section 7.4.

7.3.1 Concepts as Expectations

> ... The less we just stare at the hammer-Thing, and the more we seize hold of it and use it, the more primordial does our relationship to it become, and the more unveiledly is it encountered as that which it is – as equipment (Heidegger, 1978, p. 98).

With one metaphorical eye to the past and one to the future, *concepts are the expectations that drive experience*. Consider concepts as a tool that, once you have it, you literally cannot imagine doing without, to the extent that the tool may be incorporated into your core self-image. (See the earlier discussion in Section 4.4.3.)

Perhaps concepts are like language in this way. For many people, language is so much a part of their thinking that it seems that their thought *just is* structured as words of a (spoken or written) language. As noted in Section 3.3.2, Torey (2009) follows Davidson (1987) in making language the basis of his account of cognition: no language, no thoughts, and no mind.

Consider conceptualized experience as an emergent projection over top of non-conceptualized experience, all but obscuring it. Once we become aware of past *as* past and future *as* future, we cannot help experiencing the present moment in light of both. In Damasio's language (2000, pp. 195-233), we begin telling the narrative that gives us our rich sense of autobiographical self.

If concepts are a tool, then perhaps the metaphor is Heidegger's hammer (1978). Only when the hammer breaks or the nail bends – only when the hammer fails somehow to perform as a hammer – do we stop and see the hammer *as* a hammer. We see, hear, and feel what our concepts lead us to expect until the match between expectations and current experience breaks down in a manner that we cannot ignore. Only then are we forced to take a closer look, at which time at least some of our implicit conceptual expectations are made explicit. The unintended and seemingly paradoxical consequence is, as Heidegger notes, that we lose the "unveiled" experience of the very thing we are trying to get clear about.

This account of concepts as expectations is very reminiscent of Chrisley's Expectation-Based Architecture (EBA) (Chrisley and Parthemore, 2007b), even though Chrisley grounds those expectations (correctly, I think!) in non-conceptual content. Like the sensorimotor++ account (Section 7.2.3), the EBA is trying to combine the best elements of enactivism (in the case of the EBA, specifically Noë's version of it) and a kind of representationalism, while avoiding many of the common criticisms of both. Its key insight is to introduce "counterfactual representational vehicles": not presently tokened representations, but representations that *would* be tokened were one's (conceptual or non-conceptual) expectations to be met.

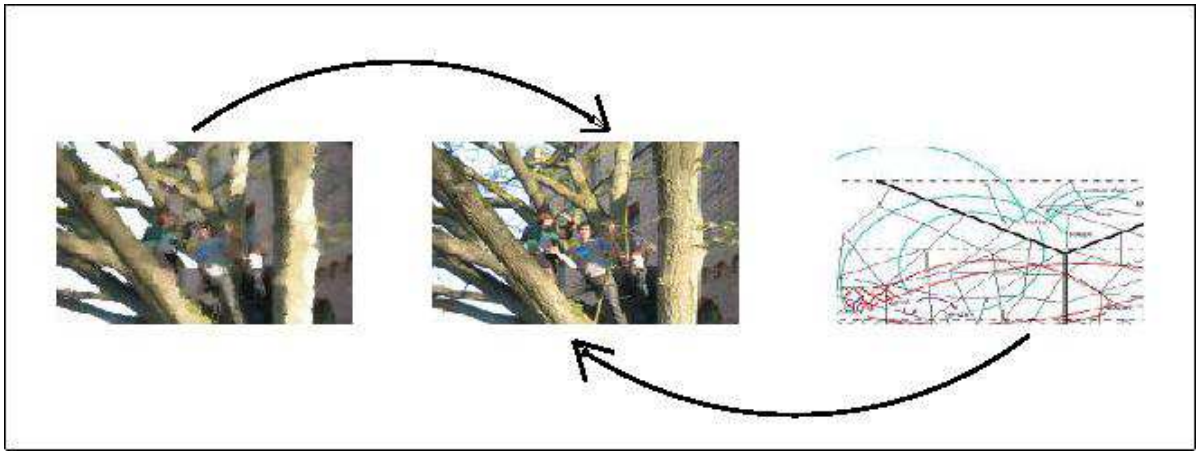

Figure 7.9: Concepts-as-representations-as-control: Consider three students in a tree (middle) and the recognition of three students in a tree (left and right). At a proto-conceptual level, that recognition will be straightforwardly picture- or painting-like (left). At a conceptual level, that recognition will be much more abstract, focusing on the logical relations of parts of the image to each other (right) - *cf.* Figure 7.8 – even while retaining links back to its proto-conceptual roots.

It is reminiscent as well of Imogen Dickie's (2006) notion of "representation as control", whereby representational (conceptual) expectations, based on past experience, guide and massively simplify the agent's interaction with its environment. (Dickie presents her own approach as preserving the best parts of the early Wittgenstein's "picture theory" (Wittgenstein, 1922). In brief, rather than the world *being* in any meaningful sense a "picture", it is rather that we represent it to ourselves and to others in a certain picture-like way. See Figure 7.9.) The problem, as Gärdenfors noted for us in Section 5.2.2, is that "...the information received by the receptors is too rich and too unstructured" (2004, p. 21).

Not surprisingly, concepts-as-expectations present several trade-offs. As noted in Chapter Five, one simplifies in order to understand. But if one over-simplifies, one no longer understands.

The set of patterns potentially discernible in any perceptual context is unbounded and may be infinite (see the discussion in Section 4.3 and (Dennett, 1991b, pp. 33-35)). Any context can be perceived from any of a bewildering variety of perspectives. Attention is limited by finite resources; there is ample psychological evidence that working memory can attend to only a small number of items at any time (somewhere around seven (Miller, 1994)).

Ask the people watching a basketball game to count, and report, the number of times the ball bounces, and they will consistently fail to see a gorilla walking across the court (a phenomenon known as *inattentional blindness*); a control group with no such instructions will be far likelier to see the gorilla (Simons and Chabris, 1999). Martin Langham reported on people who pull out in front of motorbikes, who "look but fail to see". What he found was that inexperienced drivers look all over the place. Experienced drivers minimize where they are looking (Langham, 1999).

The lesson more broadly is this: experience teaches us to become more and more selective of what we attend to. A simplified experience of the driving scene leads, *in most instances*, to improved performance. A simplified experience of the world in general may, in many instances, lead to improved performance, even better survival and reproductive opportunities. A simplified world is easier to understand and respond to. But sometimes the simplified model will make mistakes,

because the simplified model is *not* the original. Something has been lost. The driver pulls out right in front of the motorbike, even though he swears, truthfully, that he looked and saw no one.

But beyond all of this, the more conceptual knowledge we have, the more we come to rely on it. As a wealth of further psychological evidence shows, most of the time we see not what is in front of us but what we *expect* to see. Perception is not independent of reality, nor is it solely constituted by it! Instead of simply revealing the world, concepts help to construct it even as they are constructed by it.

Concepts have this dual nature: if their nature is increasingly abstract, their application is always specific to a context. To return to and refine my earlier working definition – one last time, for now!:

> A concept is (or could be described as) a synchronized pattern of relatively abstract, relatively higher-order association between some aspect of the mental world of the agent ("self") and some matching affordance(s) of her experienced environment ("non-self"), such that the affordance(s) implicitly or explicitly specifies the necessary, sufficient, and customary (or contextual) conditions for its application *relative to any particular moment*.

With this revised definition, we can begin to see how conceptual spaces theory and the unified conceptual space theory come into their own. Just as representations are neither *internal* nor *external*, being relational entities, standing between an agent and her perceived environment; just as concepts are, likewise, juxtaposed *between* the agent and her environment, created out of their dynamic interaction; so a theory of concepts properly belongs between associational and symbolic accounts of cognition, showing how conceptual mental content arises from the dynamic interaction of symbolically interpretable structures with symbol-free associations, the continuous interaction of the cognitively abstract with the sensorimotorly concrete.

7.3.2 Mapping Conceptual Space onto Perceptual Space[9]

Structuring the unified space was described as a bottom-up, layer-by-layer hierarchical (or continuous) process of *pattern recognition*: finding patterns in patterns, patterns in patterns of patterns, and so on, until the nth generation patterns have all but lost their connection back to their perceptual origins. Putting the resulting space into use is, by contrast, best understood as a top-down-driven, layer-by-layer process of *pattern matching*, a kind of "de-layering": a return down through levels of the hierarchy (or through the continuum), toward particular encounters and toward parts as opposed to wholes. If an X violates expectations, consider previous experiences with similar Xs or X-like things, or decompose the X into e.g. its functional parts. (For an illustration of this, see Section 7.3.3 and Figure 7.10.)

For concept acquisition, concepts looked more like abilities. Associations and association building were in the driver's seat. For concept application, concepts look more like representations. Initially, at least, symbols and symbol application are the more appropriate level of description (though only some part of this need be consciously articulable by the conceptual agent herself).

[9] It may be useful to compare the discussion here with that of Section 7.2.5.

The basic idea is this: unified space can be fitted to perceptual space, concepts in the unified space matched against their non-conceptual analogues in present experience, attempting the closest match possible: a *particular* concept (most likely at the basic-level on Rosch's hierarchy – see (Rosch, 1975, p. 587)) to a *particular* sub-region of present experience. A precise match may push one to further identify the match at the subordinate level. A mismatch, on the other hand, may force one to jump to some *other* basic-level category entirely. (That "dog" in the shadows is not a dog at all; it's a rock. That "wailing cry" was not a woman in distress; it was the call of a loon.) Alternatively, one might, as it were, relax the resolution: resorting to superordinate categories that match against larger and larger portions of the unified space, going more and more general until a match finally does occur. (At some point a match *must* occur, since everything perceivable is, minimally, a *something*.)

So: sometimes a mismatch can be resolved by jumping to another location in the unified space. Other times it can be resolved by "zooming out" and in again. Other times, however, there will be breakdowns: failures of our conceptual expectations. Breakdowns suggest four strategies for dealing with them, which we can order from least to most radical[10].

1. Adjust the logical structure of the closest matching concept (Section 6.2.2) in terms of its parameters, contextuals, and, if appropriate, components, so that a match now does occur. Formerly you conceptualized all swans as white; now you conceptualize swans as either white or black.

2. Alternatively, perhaps you did not conceptualize swans in terms of colour at all, but now you must, in order to distinguish the white from the black swans. This is the equivalent, in terms of traditional conceptual spaces theory, of adding an additional integral dimension (namely *colour*).

3. Partition an as-yet-unpartitioned sector of the conceptual space. Formerly you recognized only a general category of swans, all of which were white. Now you recognize two distinct sub-categories of swans, one of which includes all the white swans and one all the black swans.

4. Most radically, remove partitioning from some sector of the conceptual space and re-partition it: i.e., divide it into a different set of (convex) shapes. As opposed to conceptual change via (1) or (2), this is conceptual obsolescence and replacement. This would be if your encounter with black swans forced you to e.g. re-structure your whole "bird" space, classifying all birds differently and not only swans.

[10] *Cf.* the discussion in (Aisbett and Gibbon, 2001, p. 210)

7.3.3 On Encountering a Door (A Thought Experiment)

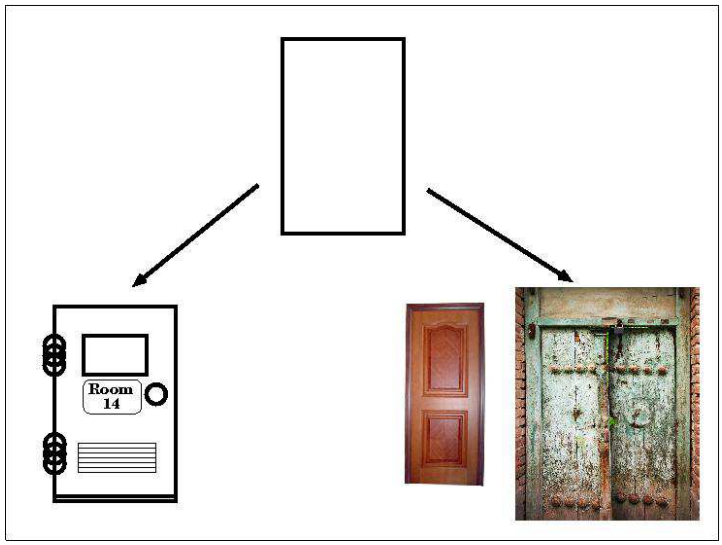

Figure 7.10: A door may be taken as an undifferentiated whole, a set of functional parts, or a similar instance to some previous door encounter.

Consider a door that is in front of you (see Figure 7.10). Does your present experience of that door *as* a door match your expectations at the most abstract conceptual levels with respect to doors? If you don't need to see the door as anything more than a whole with no parts (like an "unstructured" symbol), then you won't: it will register as an undifferentiated door (top of figure).

Of course, depending on where your attention is focused, you may choose or be motivated to look more closely, by considering the door as a special *type* of door (like a trap door, or the kind of door one finds on a bank vault). Alternatively, you might focus on its functional parts (lower left): where is the handle?... where are the hinges?... does the door have a sign on it? Again, you might consider the door in the light of particular past door experiences (lower right). Does this door remind you in some way of one of them?

How you encounter the door will depend, in part, on what you want to do with it. If you need to pass through the door, you will look, minimally, for how the door opens. If it has a handle, you'll probably be inclined to pull it. If it has a flat metal plate where the handle would be, you'll be inclined to push it. The more closely you examine the door, the more directly your sensorimotor capacities with respect to that or other doors will be brought to bear, on-line (directly engaging with the door) or off-line (simulating that engagement).

Only if the door has something perceptually un-door-like about it will you be forced to examine it yet more closely: e.g., if the door has a handle but is meant to be pushed instead of pulled. One could imagine that the "door" is only a painting on the wall, or has been painted or nailed shut. Unusual doors will focus your attention and shift it from the abstract and general to the concrete and immediate, from doors as some platonic-like entities to specific door encounters.

In the process, you may come to change your understanding of doors a little; or you may derive a concept of a new kind of door. You might re-structure your entire "door" space. Of course, at some

point, the unusual door in front of you may confound all attempts at conceptual understanding, and you may resort to brute sensorimotor engagement with it!

7.4 Which Takes Precedence? (Some Final Thoughts on Representations)

The ordering of sections 7.2 and 7.3 was not, as noted at the start of the chapter, purely arbitrary. Application preceding acquisition seems intuitively as impossible as effect preceding cause or – for the empiricists at least (of which I admit to being one) – understanding preceding experience.

7.4.1 Locating the Observer

Furthermore, as much as I owe a debt to the phenomenologists, my background is rather from analytic philosophy, which strives to be a scientific discipline. As noted in Section 5.2.1.1, empirical science has accomplished a great deal by pushing the observer into the background or ignoring the observer's role altogether. The story I have told of concept acquisition is the one that has likewise backgrounded the observer; while the story of concept acquisition has brought the observer back into the foreground. So in telling the acquisition story first, I am exposing my analytic roots. At the same time, recall the conclusion I made in the final section to Chapter Five:

> . . . Sciences of concepts and of consciousness remind us of what we should have remembered all along: that the observer is *always* there. . . that the subjective is inseparably bound up with the objective, that science yields up not timeless understandings freed from cultural and historical contexts but working hypotheses.

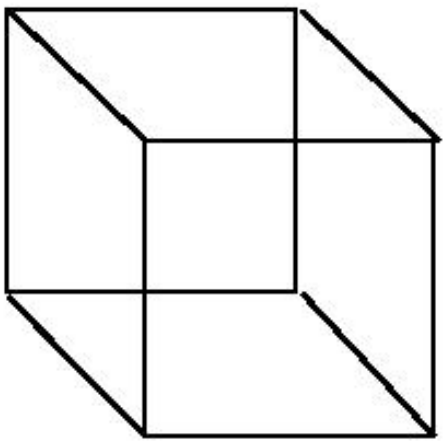

Figure 7.11: Necker cube: Does it extend from or recede into the page?

Often the backgrounding of the observer is a good strategy, perhaps even the *right* strategy – but not always. Sometimes, the order of things needs to be turned around. Sometimes we require a contrasting perspective; and sometimes our perspective inverts of its own accord, as when we view a Necker cube (see Figure 7.11), or when we look at the famous illustration by W.E. Hill of

Figure 7.12: Hill's young woman / old woman, circa 1915: it is possible to see one or the other, but not both at the same time. (Photo downloaded from Wikimedia Commons: http://commons.wikimedia.org/.)

the young woman / old woman (see Figure 7.12). When we are discussing representations, the observer is (or should be, on my account) very strongly in the foreground.

7.4.2 The Wider Debate Over Cognition

Earlier I noted that, in my concept acquisition story, concepts looked more like (non-representational) abilities; while in my concept application story, they looked more like (what some insist on calling mental) representations. I introduced the debate between concepts as abilities and concepts as representations in Chapter Two. That debate found its clearest expression in the table toward the end of Chapter Five, where I suggested that most contemporary theories of concepts, and most debates about them, line up on one side or the other of the divide. The idea of these two competing, never-truly-resolvable perspectives has been a running theme throughout this work.

It is time to put the debate between concepts-as-abilities and concepts-as-representations into the context of a wider debate over the nature of cognition, and in doing so bring it to a close. That wider debate has been lurking in the background through all the preceding chapters. It is one that is longstanding, often bitter, and so far unresolved; taking many forms and couched in various terms: pitting representationalists against anti-representationalists, cognitivists against connectionists, symbolists against associationists, and rationalists against empiricists (*cf*. (Brooks, 1991a,b; Chalmers, 1990; Fodor and Pylyshyn, 1988; Perry, 1986)). One way to frame the debate is this: is mental content *of any sort* best understood in terms of symbols or symbol manipulation (where the brain is thought of as e.g. a physical instantiation of an abstract Turing machine), or in terms of explicitly non-representational associations between various observed regularities (on

different levels of concreteness and abstractness)?

Of course, those on the one side of the aisle do not deny that associations play a role. But the meaning is in the things being associated, not in the associations; and context is deemed, at least most of the time, largely irrelevant. Likewise, those on the other side of the aisle do not deny that we employ symbols, nor (necessarily) insist that "representation" is an empty term; but for them, the perspective is turned around: meaning is in the associations, not in the things being associated. Context is key, and local details of structure are largely irrelevant. Push either side hard enough, and some ground will be ceded to the other; but on both sides, the tendency is to think that one side or the other *must* be right.

7.4.3 The Death of Representations

If I may borrow a line from Mark Twain, the death of representations has been greatly exaggerated. Concepts are, indeed, more than representations, logically; but, at the same time, representations are *unavoidably, for us*, part of what it is for something to be a concept: for when an agent reflects upon some concept (or on concepts in general), or otherwise employs a concept reflectively, she is using it to represent *something* to *someone*, be it herself or another agent. Representations, properly understood, are both *necessary* for understanding cognition, *even at the most basic sensorimotor levels*, due to our inability as reflective conceptual agents to step aside from our representational perspective; and *not sufficient* for understanding even the most abstract levels of symbolically structured thought, given the logically ubiquitous involvement of (non-representational) sensorimotor engagements throughout cognition. The representations need not, in any sense, be in the mind of the agent being observed, but rather are bound to the perspective of the agent doing the observation (who nonetheless may project it into the mind of the agent observed!).

At the same time, representations are not just some awkward necessity; they have real value. In responding to Rodney Brooks (1991b), David Kirsh talks of concepts, but it is representations, I believe, that are his ultimate target:

> There is a limit... to how far a creature without concepts can go.... Concepts are either necessary for certain types of perception, learning, and control, or they make those processes computationally simpler. Once a creature has concepts its capacities are vastly multiplied (1991, p. 191).

What representations gain us, argues Richard Shusterman in response to anti-representationalist Merleau-Ponty, is the capacity for explicit reflection and the consequent ability to recognize and modify habits, including bad ones:

> ...In order to effect... improvement, the unreflective action or habit must be brought into conscious critical reflection (if only for a limited time) so that it can be grasped and worked on more precisely (2008, p. 63).

He could as well be responding to Rodney Brooks when he says, "the claim that we can do something effectively without explicit or representational consciousness does not imply that we cannot also do it with such consciousness and that such consciousness cannot improve our performance" (2008, p. 68).

So which takes precedence: concepts or experience, acquisition or application, reason or empirical discovery, action or thought? It all depends on where you stand, and how you look. Partly it is a matter of which questions you are asking, partly on the context in which you intend to apply the answers; but the tension and the oscillation between competing views is, I believe – in contrast to a similar oscillation McDowell addresses and believes himself to resolve in ***Mind and World*** (1996) – eternal[11].

7.5 Conclusions

How do concepts (as abstract and structured entities) and experience (seemingly immediate, and largely unstructured) inter-relate? That has been the focusing question of this chapter, as I have sought to move the discussion out of the largely static view of concepts from the earlier chapters toward a dynamic view of continuous acquisition and application. In the process I have borrowed a page from the classical definitionists, for whom definitions were neutral as to whether they were defining new concepts or verifying existing ones. Indeed, the final (though still very provisional!) definition of concepts that I offer owes quite a lot to those early traditions, even while the spirit of my approach often sides with the prototype and similarity-space-based accounts, which derive more from the imagist traditions.

Concepts *do* have something peculiarly definition-like about them, once the requirements that definitions be strictly static and strictly public entities are relaxed. The appropriate metaphor here might be that of a "dynamic dictionary", where the words on the page are constantly in motion: look at a definition, look away, look again, and the definition has subtly changed. Offering a definition in the more usual, dictionary-type sense of the word becomes an intellectualized attempt to fix the concept, to take a snapshot. Something of what the concept is gets captured (hopefully), but something more gets lost.

Concepts without the capacity for motion, without the capacity for change, are dead. The same might be said of any theory about them.

The other primary motive of this chapter has been to put the ideas about conceptual spaces and the unified conceptual space to practical work. I showed how, in my account of concept acquisition through ontogenetic development, an initially unstructured conceptual space becomes increasingly partitioned and topologically complex. At the same time as proximal points in the unified space take on increasing significance, so do distal points, as parts of that space are mapped onto each other.

Turning the acquisition process on its head, concepts become *the expectations that drive experience*. Conceptual space is mapped back onto the perceptual space from which it has been derived, layer by layer, until matching succeeds or breakdown occurs. Breakdown can prompt a variety of responses, from the conservative (tweaking the concept's logical structure) to the radical (re-partitioning a particular conceptual space).

[11] McDowell's oscillation is between a kind of anti-realist coherentism – beliefs are justified because they all support and are consistent with each other – and justificatory appeals to a non-conceptual Given (what he terms the *Myth of the Given*). See in particular Lecture One, pp. 3-23. My own approach in this work has been to deny, with McDowell, that myth; while also denying that either experience or world are fully conceptualized.

Central concepts in this chapter have been:

- *Dynamically coupled systems*, which are, logically speaking, impossible to separate, even while it may be conceptually useful or even necessary to (attempt to) do so.
- *Circular causality*, by which effects can be seen, from some perspective, to be their own causes, like the dragon eating its own tail (Figure 7.2).
- *Co-emergence*, a more conceptually abstract way of talking about certain dynamically coupled systems, whereby two opposite or opposing forces can be seen each to give rise to the other, like *yin* and *yang*.
- *Sensorimotor engagement*, whereby the most abstract levels of cognition are grounded in basic agent-environment interactions, in which the sensory and motor systems are dynamically coupled, while sensory experience and motor action are co-emergent.

The chapter's principle conclusion is that concepts and experience are co-emergent within a causally circular loop of sensorimotor-grounded conceptual acquisition and conceptual-expectation-driven application, where neither has priority over the other.

As a coda, I tied the debate over concepts-as-abilities versus concepts-as-representations back into a wider, often acrimonious debate over the nature of cognition: is cognition *itself* to be understood primarily in representational or non-representational terms? The "correct" answer depends, once again, on your vantage point.

Chapter 8

From Theory to Practice: A Simple Application

As noted back in Chapter One, the present research began as a project in cognitive science, attempting to address what I and others (notably Fodor (1998)) have perceived as a chronic shortcoming in the field: a failure to be sufficiently explicit about the theory of concepts being employed in any given research project, on the assumption that *all* research in cognitive science depends, in the end, on having a theory of concepts and not getting it *too* wrong. Where the theory of concepts is left implicit, there is, by definition, no attempt to justify that theory philosophically, psychologically, biologically (including neuroscientifically), or otherwise.

When one shifts from theoretical cognitive science to applied artificial intelligence (AI), the problem becomes all the more stark: knowledge, including explicitly conceptual knowledge, is represented in the software applications in whatever way *works.* Certainly, the capturing of "conceptual" knowledge in e.g. XML (*extensible markup language*) is difficult to justify in any other way. This need not be a problem – depending on one's goals. There is often much to be said for quick and dirty.

On the other hand one might believe, as I do, that AI, as the engineering arm of cognitive science, has much to gain from using models of concepts and conceptual content that are as isomorphic as possible (on one or more dimensions) to the concepts and conceptual content we appear to know the best: our own. In that case, even the quick-and-dirty approaches stand to benefit; and for those whose goal is not merely to emulate one aspect or another of intelligent behaviour but to capture something of the holistic essence of intelligence itself – if there is such a unitary thing – such considerations may be crucial.

Despite its grounding in cognitive science, the discussion to this point has taken place almost entirely within the relatively rarefied domain of philosophy of mind (and there within the even more rarefied sub-domain of theories of concepts). I have made mention of biology, neuroscience, and psychology (as well as mathematics) where I felt sufficiently well informed to make a meaningful comment.

On the one hand, if my earlier arguments have any merit, then the lines we draw between disciplines, like the conceptual boundaries we draw anywhere, are pragmatic and ultimately arbitrary, masking what is, from a conceptual point of view, an underlying continuum. All disciplines are, to some

unavoidable greater or lesser extent, interdisciplinary. If cognitive science claims to be special this way, it is only because of how it *actively promotes* the crossing of disciplinary boundaries.

On the other hand and by those same arguments, we *must* draw lines, both from necessity – it is in the basic nature of concepts to do this – and from practicality. As I suggested in Chapter Four, a theory of concepts that tries to do everything does nothing well. Worse, if the arguments from Chapter Five are correct, such a theory will slide into vicious circularity, yielding either incoherence or blatant inconsistency. Better to restrict one's domain of interest and of application, even if that restriction is somehow arbitrary.

So, again, my goal has not been to set out a complete theory of concepts, nor to claim that my account is fully consistent. Indeed, in terms of the relationship between representations and (non-representational) abilities (Chapter Two), the "toggling effect" it serves as an example of (Chapter Five), or the relationship between concepts and experience (Chapter Seven), that account actively exploits certain unresolvable inconsistencies.

My goal, rather more modestly, has been to lay the foundations from philosophy of mind for a new approach to concepts in cognitive science (in general) and in AI (in particular) – by assessing, reconciling (where possible), and otherwise synthesizing the various approaches. If I have favoured some approaches over others, it is because they are, I believe, better suited to assist this synthesis.

In contrast to the earlier chapters, this chapter will be much less philosophy of mind and much more cognitive science, less theoretical foundation and more application. It poses the question: how might the theory of concepts outlined in these chapters be used to create a concrete software application, and how could such an application be used to motivate empirical research?

If my arguments on the ineliminability of the observer are correct (Section 2.8), one of the consequences is that there need be no ultimate objectivity, if by that one means an objectivity divorced from subjectivity. There may only be a practical sort: a distancing or "stepping back" from the limitations of *individual* perspective. Perhaps, though, that is all the objectivity we need; and at some future point, one might deign to hope that any lingering pretensions to a platonic realm will be laid to rest. Not coincidentally, I have described concepts as doing a very similar kind of "stepping back": helping us to understand at the same time, ironically, that they distance us from the very things we are trying to understand – enabling us with the one hand, constraining us with the other.

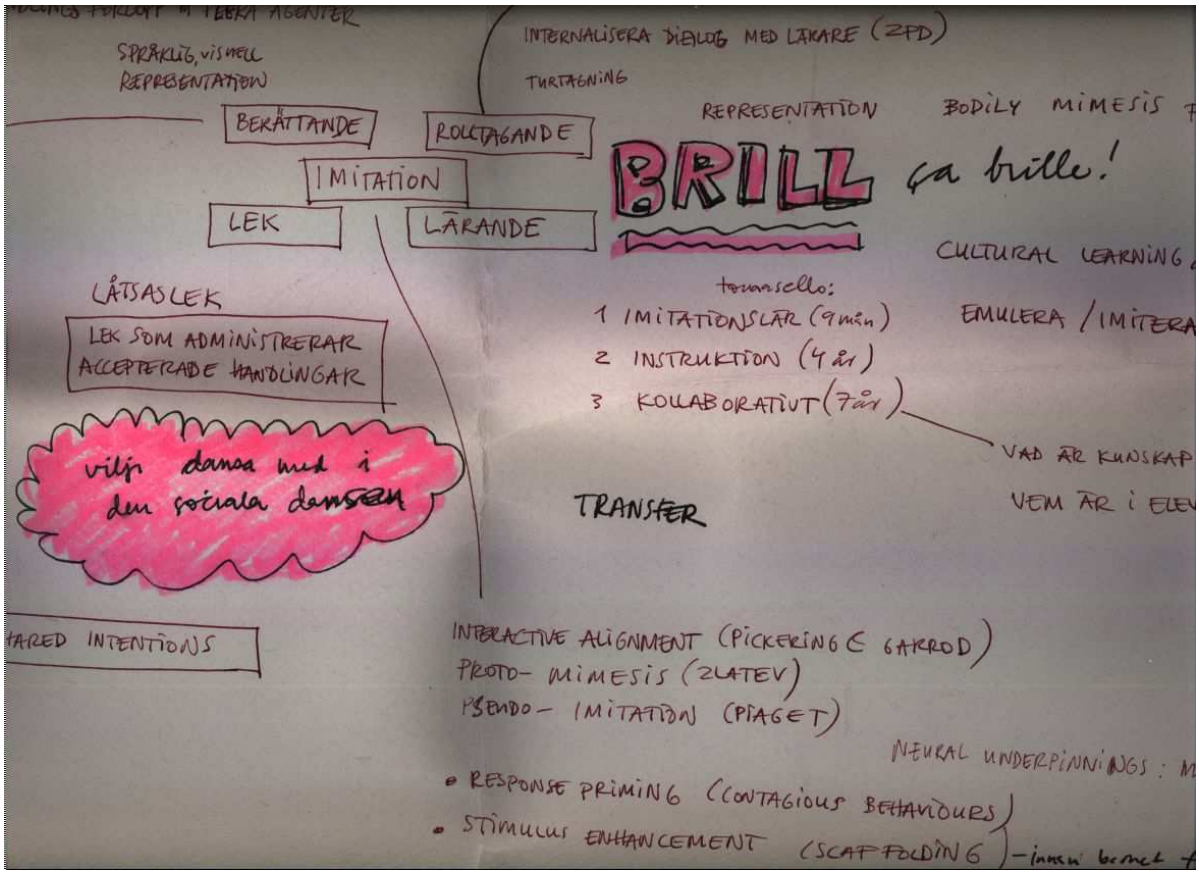

Figure 8.1: A portion of a hand-drawn mind map in Swedish and English, copyright Åsa Harvard. Used by permission.

The toy application described below is meant to facilitate that practical kind of objectivity, by offering the user a better set of tools with which to construct an externalized model of some portion of her conceptual domain: mapping out and then exploring the territory. What is externalized is thereby made explicit; what is made explicit is easier to scrutinize; what is easier to scrutinize is *ceteris paribus* easier to modify.

Figure 8.1 offers an example of a hand-drawn mind map, sometimes called a concept map. (For possible distinctions between the two, see Section 8.1.2.) Mind-mapping software – such as the toy application I describe – mimics and formalizes (hence constrains) the process described in the figure. What makes the toy application different is that it implements a specific theory of concepts. More particularly, it implements many of the features of the unified conceptual space theory presented in Chapter Six and explores the "application" side of the acquisition/application model offered in Chapter Seven.

First, though, I survey the existing field of mind-mapping software, with particular attention to some of its present limitations.

8.1 Existing Mind-Mapping Software: A Survey

A significant number of free and commercial packages exist, all with essentially the same functionality; the principal differences lie in sophistication of user interface. Nodes radiate outward from an original parent node, linked in one of two ways: either by a parent-child link or by a cross-reference

link (which can link two nodes anywhere in the map). One has a limited choice of shapes for the nodes and of styles for the parent-child links. Individual nodes can be annotated with icons, images, notes of arbitrary length, hyperlinks, and, in the case of the commercial software, arbitrary attachments. Nodes and links can all be colour-coded. At any node, other, previously existing maps can be added as sub-maps. Large maps can be scrolled or zoomed in or out. Maps can be exported as linearized text to e.g. a word processor file.

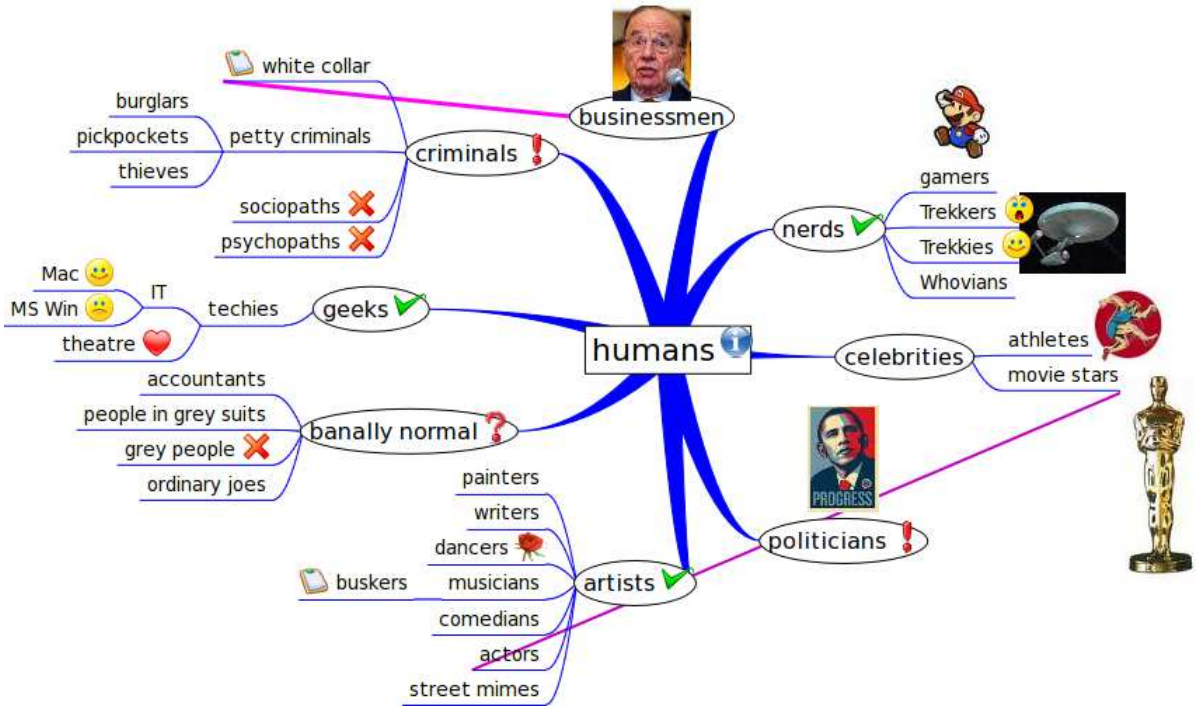

Figure 8.2: VYM screenshot: a mind map for "humans".

Figure 8.2 shows a screenshot from VYM (View Your Mind)[1], describing the author's conceptual space of "humans". In this case the blue lines indicate the parent-child links and the pink lines the cross-reference links. A "notebook" icon indicates an attached note (i.e., adjacent to "white collar" and "buskers").

8.1.1 Theoretical Justification

The central driving intuition behind mind-mapping software is, first, that underneath the apparently linear, apparently (at least to many people) propositionally structured nature of self-conscious, directly introspectible thought lies a much more richly structured layer that is non-linear and non-propositional. Such a view I take to be untendentious, and is expressed well by e.g. the psychologist and neuroscientist Joseph LeDoux:

> The conscious and unconscious aspects of thought are sometimes described in terms of parallel functions. Consciousness seems to do things serially, more or less one at a time, whereas the unconscious mind, being composed of many different systems, seems to work more or less in parallel. Some cognitive scientists have suggested that consciousness involves a limited-capacity serial processor that sits at the top of the

[1] Available from http://www.insilmaril.de/vym/.

> cognitive hierarchy above a variety of special-purpose processors that are organized in parallel... (1996, p. 280).

Second, software tools that directly support that largely if not entirely unconscious level of cognition may assist e.g. the writing process in ways that tools aimed at the conscious/linear/propositional level may not. In particular, they should support brainstorming by freeing the user from having to structure her ideas into a linear text, and they should allow her to organize her ideas in a visually much richer way than such a text would allow. One person who has spent considerable time exploring these possibilities with respect to so-called "writing environments" is Nottingham University's Mike Sharples; see for example the discussion in (Sharples, 1999, pp. 77-82).

In an educational context, mind-mapping software can assist either student or teacher in evaluating the student's understanding of a particular domain (i.e., examining it for completeness and consistency) (Novak and Canas, 2008). Skeleton maps can serve as teaching aids, where a "skeleton map" is one that "has been previously prepared by an expert on the topic, and permits both students and teachers to build their knowledge on a solid foundation" (Novak and Canas, 2008).

8.1.2 Related Concepts

Concept maps overlap substantially with mind maps; however, mind maps are typically more structured and are organized around a single central idea. Concept maps are typically more freeform and may have multiple (albeit related) centres. As opposed to the standardized parent-child links in mind maps, concept maps may have an unbounded number of different kinds of links, each with its own label. Figure 8.3 shows a screenshot from CmapTools[2]. As before, a notebook icon indicates an attached note. Notice how the presentation is less graphically rich and less tree-like: the emphasis is on a diversity of nodes and links.

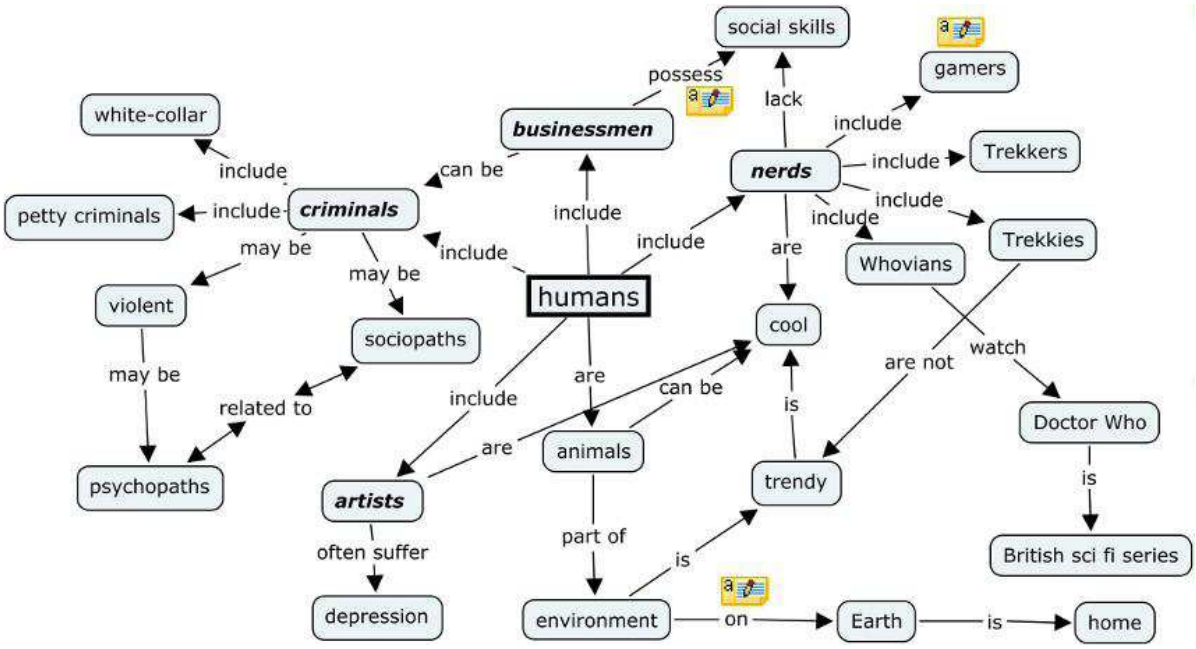

Figure 8.3: CmapTools screenshot: a concept map for "what sort of people are humans?".

[2] Available from http://cmap.ihmc.us/.

This survey is in no way intended to be complete. One can find many more variations, including e.g. argument maps (similar to what Sharples (1999, p. 80) calls "notes networks"), which are meant to help explicate arguments, either for the people trying to build them or the readers trying to understand them.

8.1.3 Limitations

> Mind maps have been invested with almost miraculous powers (Sharples, 1999, p. 80).

The main drawback to all of these packages is that, while they are indeed based on certain broad premises about cognition, an explicit and comprehensive theory of concepts, or of cognition more generally, is lacking. (Maria Ruiz-Primo and Richard Shavelson say, rather more graciously, that "... there is a cognitive theoretic basis for concept maps, but the variation observed in the use of concepts maps seems to be unrelated to this body of theory" (1996, p. 573).)

The nature of concepts is in some way presupposed; the maps then allow the user to connect nodes together into higher-order structures. There is something black-box-like about the nodes (or, in the case of concept maps, the nodes and the links). That might be fine if one adopts Jerry Fodor's informational atomism (see Section 2.4.1), under which concepts have no conceptually relevant substructure; but I have argued consistently throughout this book against such a position.

To be clear: I am not saying that these packages lack theoretical foundations altogether, only that they lack certain foundations which, from the perspective of the present work, are critical, and that their proponents might do well to acquaint themselves more with the literature on concepts in philosophy of mind. In consequence of these shortcomings, mind maps and concept maps are, by any concept-sensitive account, massively under-constrained. There is no one technique either on how to construct or evaluate them. "Unfortunately we cannot look to cognitive theory to decide which technique to prefer, because many of the techniques [we] reviewed had no direct connection with such a theory" (Ruiz-Primo and Shavelson, 1996, p. 585). What makes a good or a bad map is determined not by the rules for its construction, which are sparse; rather, the determination is *post hoc* and, worryingly, largely *ad hoc*.

Ruiz-Primo and Shavelson are concerned with the use of concept maps as a form of assessment in science education, and their criticisms are targeted at such usage. However, I think their criticisms raise far more general concerns. They note, for example, that while studies do indeed show that novice maps can reliably be distinguished from expert maps for any given domain, "... results also suggest that experts' concept maps are not as similar as we would want and expect" (1996, p. 593). Partly this may have to do with the under-constraining of the maps; partly it may relate to the public/private distinction I first raised with respect to concepts back in Section 3.3.3. These maps may, as well, be predisposed to capturing some kinds of knowledge over others: "... concept maps are more directly related to the knowledge of facts and concepts and how the concepts in a domain are related than how they are used or applied to solve a problem" (1996, p. 573).

On these, as well as many other conceptual matters, the mind and concept map designers are largely silent.

8.2 Charley: A New Kind of Mind Mapping

I must be clear at the outset that the application described below is nothing more, at present, than a proof of concept. At the same time, it suggests a radically different approach to mind mapping, and allows us to see the way forward to empirical testing in the concluding chapter.

8.2.1 Implementation

Charley is a sophisticated Voronoi tessellation generator and navigator. It is written in TCL/TK: i.e., a combination of the high-level scripting language TCL ("tool command language") with the TK graphical toolkit (set of widgets), commonly used for rapid prototyping. It interfaces with a program by Steven Fortune, written in C, which implements Fortune's (1987) "minesweeping" algorithm for generating Voronoi diagrams. That program inputs a file containing a series of points as *x, y* coordinates:

```
205 413
428 410
346 548
```

. . . And generates this sort of output:

```
s 428.000000 410.000000
s 205.000000 413.000000
l 1.000000 -0.013453 310.964111
s 346.000000 548.000000
l -0.594203 1.000000 249.043472
v 316.847260 437.315033
l 1.000000 0.957447 735.553162
e 0 -1 0
e 2 -1 0
e 1 0 -1
```

. . . Where "s" indicates a node at $\boldsymbol{x}$, $\boldsymbol{y}$; "l" indicates a line of form $\mathbf{a}x + \mathbf{b}y = \mathbf{c}$; "v" indicates a vertex at $\boldsymbol{x}$, $\boldsymbol{y}$; and "e" indicates a line segment (as part of line $\boldsymbol{m}$) attaching to vertex $\boldsymbol{n}$ or extending to infinity (-1). The TCL/TK program converts this to graphical format and renders it on screen.

8.2.2 Operation

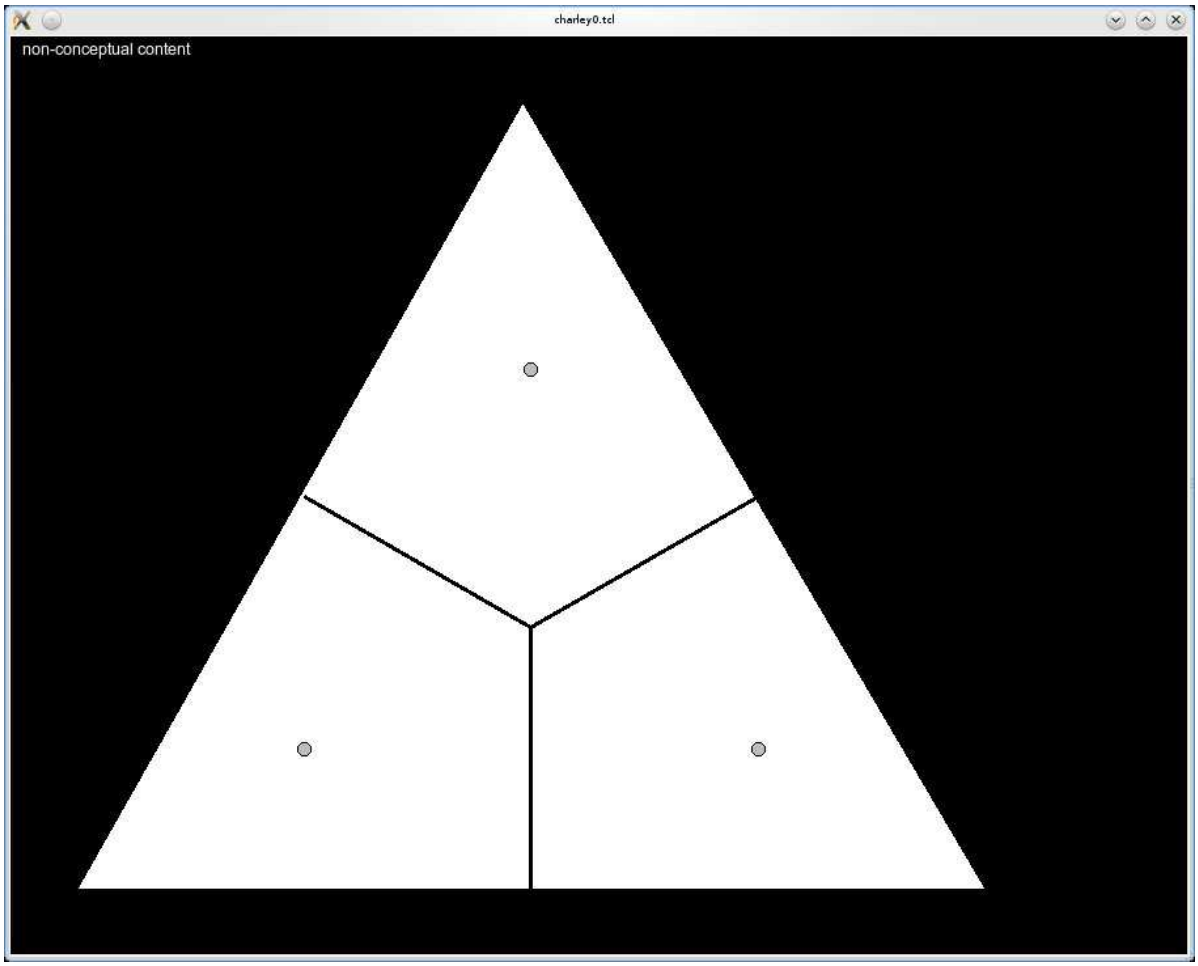

Figure 8.4: The initial window showing the space of concepts (white) within the surrounding space of non-conceptual mental content (black).

Figure 8.4 shows the program's start-up screen. The white triangle represents "the space of concepts" (the top-level concept in the concept hierarchy) whilst the surrounding black area is the neighbouring "non-conceptual content". Shifting the mouse between the white (current) and black (neighbouring) areas changes the label at upper left accordingly. Within the space of concepts are three sub-regions: "the space of objects" (at left), "the space of action/events" (at right) and "the space of properties" (at top). Hovering over the prototype-point (grey circle) for the sub-region highlights the circle (it becomes blue) and, again, changes the label at upper left accordingly.

The program implements the following algorithm:

1. Left click anywhere in an unpartitioned sub-region to create a new node (Figure 8.5)[3].
2. Left click anywhere else in the same sub-region to create a new node and re-partition (re-tessellate) the sub-region (figures 8.6 and 8.8).
3. Left click and drag any existing node to drag it to a new location (Figure 8.7).
4. Right click in any sub-region to attach a new label to that sub-region (Figure 8.9).

[3]Compare this list item for item to the list in Section 8.2.3.

5. Right click on any point to link to a new point or jump to an existing one in another region of the space (Figure 8.10).

6. Scroll the mouse wheel to zoom in or out (Figure 8.11). (To zoom in, the mouse must be over the prototype-point. To zoom out, the mouse can be anywhere in the currently highlighted sub-region.)

7. Left click on neighbouring sub-regions (in black) to scroll left-right or up-down (Figure 8.12).

Note that (1), (2), (3), and (7) apply only below the top level in the concept hierarchy, while (4) and (5) apply only below the top *two* levels (i.e., one cannot re-name a proto-concept nor change its relations to the other proto-concepts). Meanwhile for (6), one cannot (of course) zoom out beyond the top level.

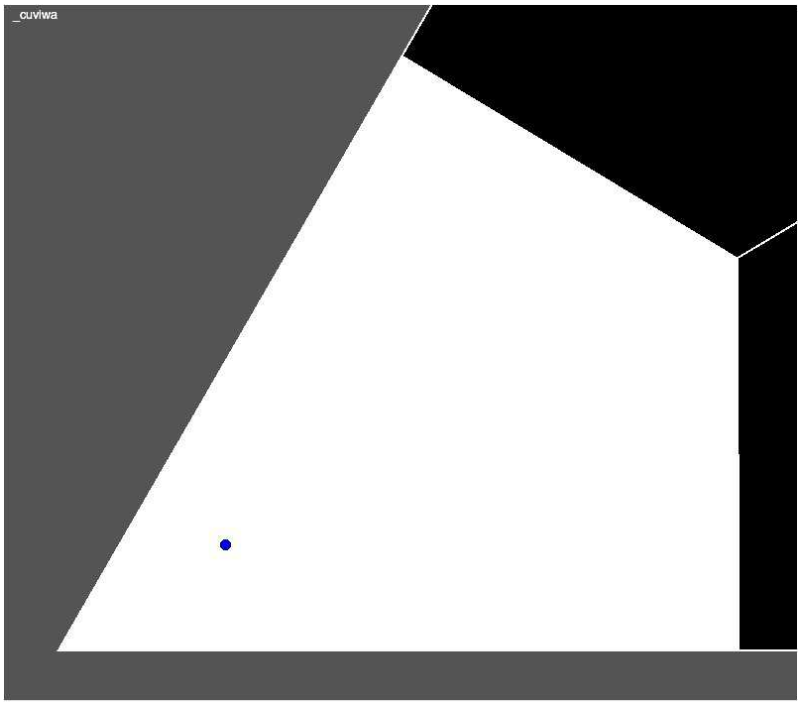

Figure 8.5: Left click anywhere in the active sub-region (white area) to create a node one layer down in the hierarchy. The node is initially assigned a randomly generated name, which is displayed when the cursor is over the node. Otherwise the name of the parent node is displayed – in this case, the "space of objects".

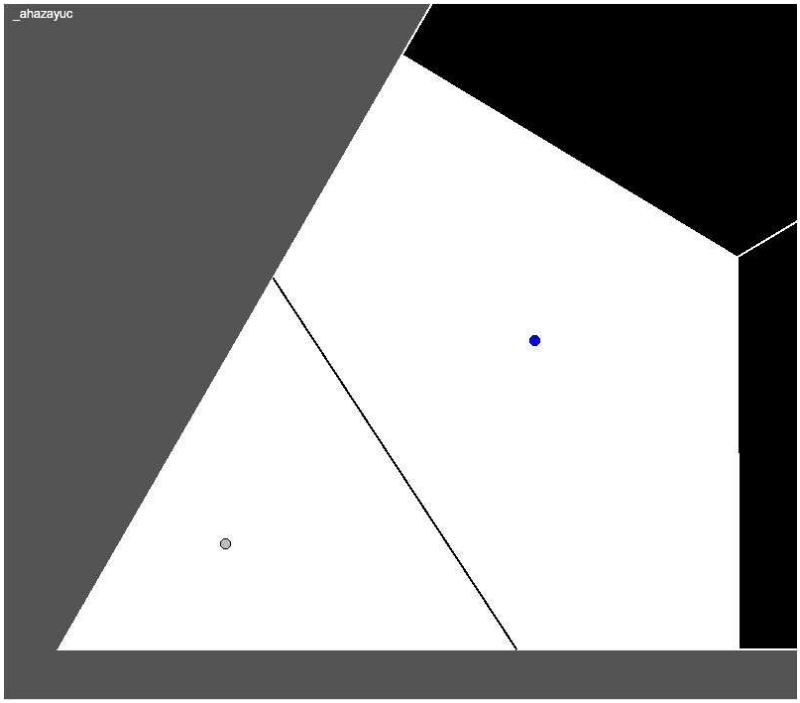

Figure 8.6: Left click again to create a new node. Two or more nodes forces a (re-)tessellation.

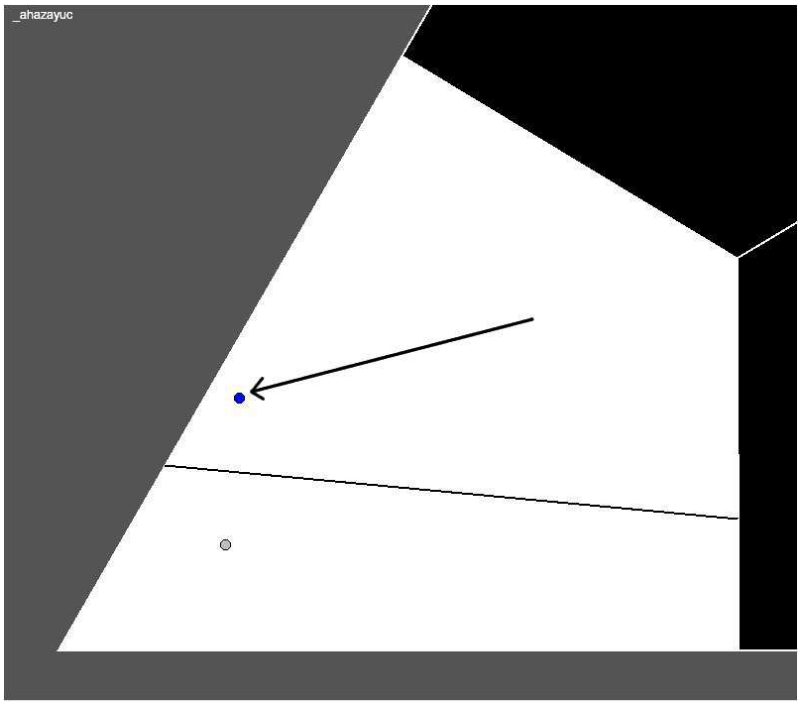

Figure 8.7: Left click and drag a point to force a re-tessellation.

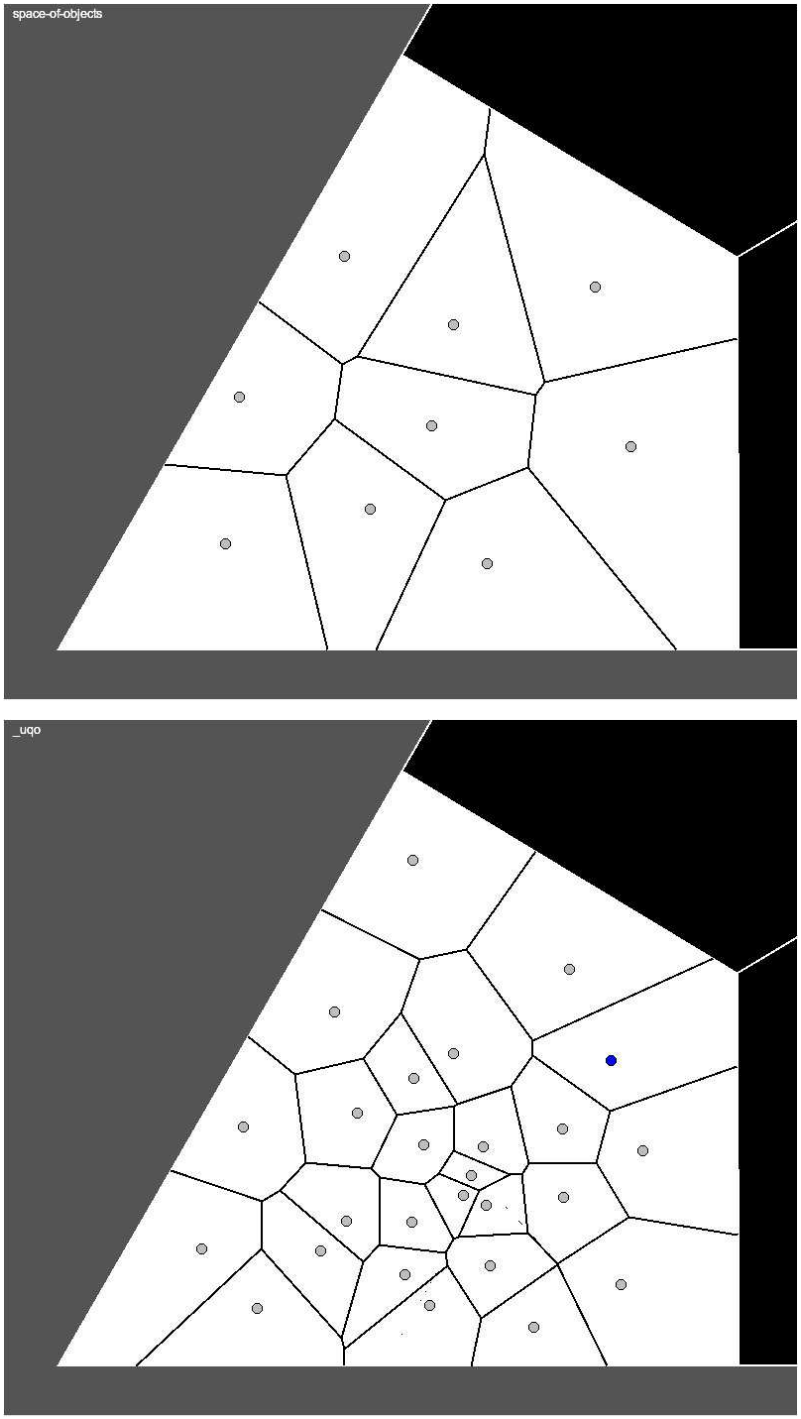

Figure 8.8: The same sub-region, after several additional nodes have been added.

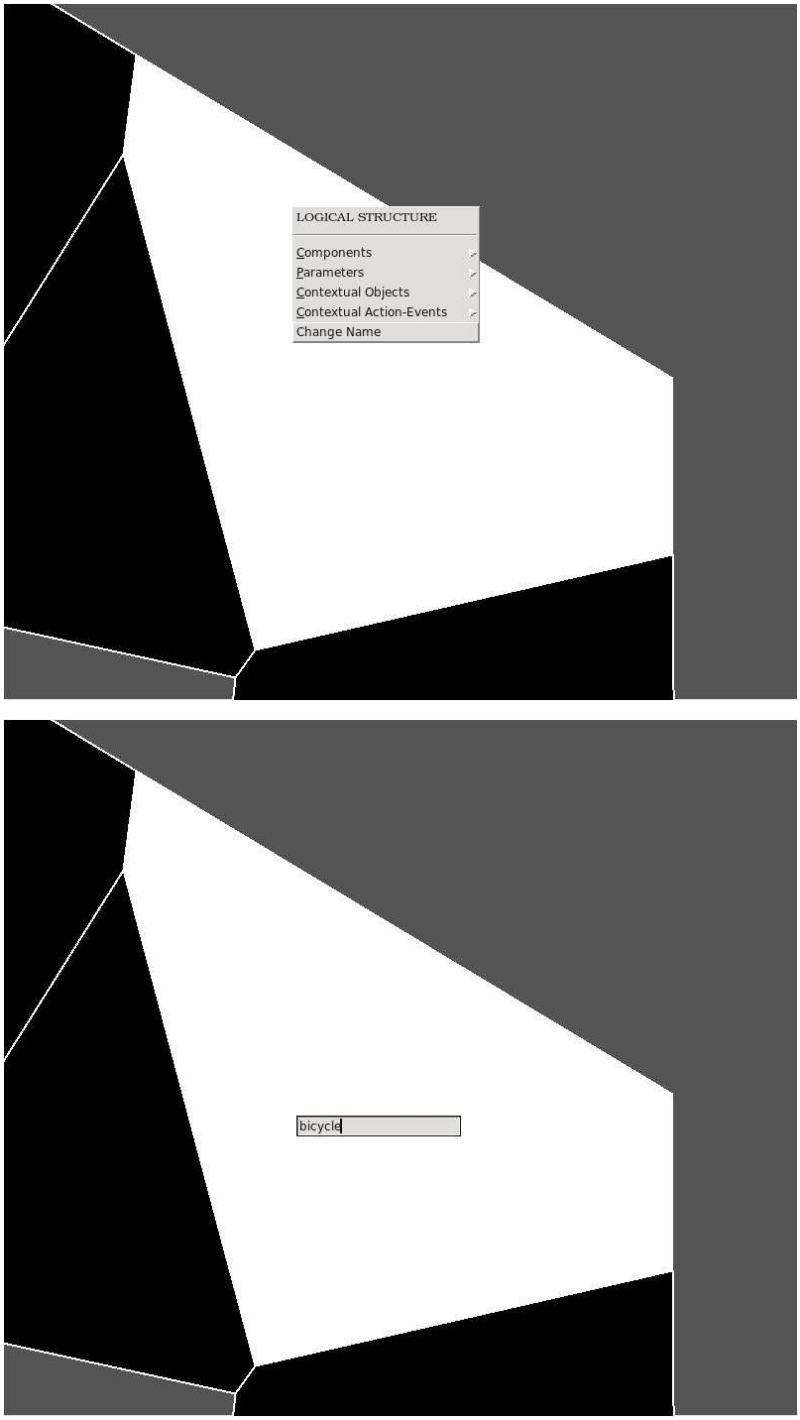

Figure 8.9: Right click to change the label on any sub-region: the "Change Name" entry is automatically highlighted. Clicking on it brings up a simple dialogue box.

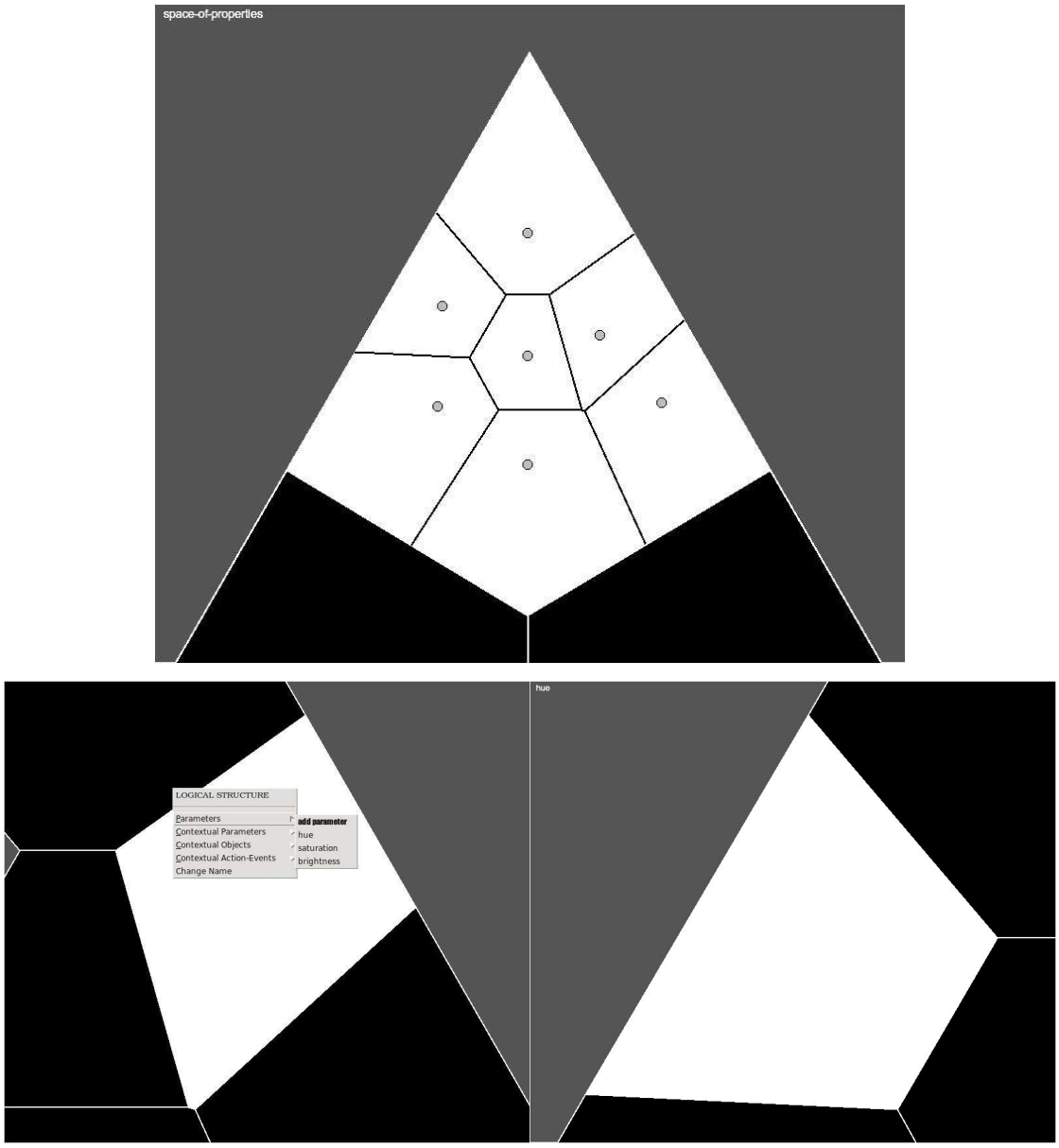

Figure 8.10: Adding and moving between links: the "colour" node (lower left, and at right in the top picture) has the parameters (integral dimensions) of "hue", "saturation", and "brightness". Clicking on "add parameter" will bring up a dialogue box for adding another parameter. Clicking on "hue" shifts the focus to the "hue" node (lower right, and at left in the top picture).

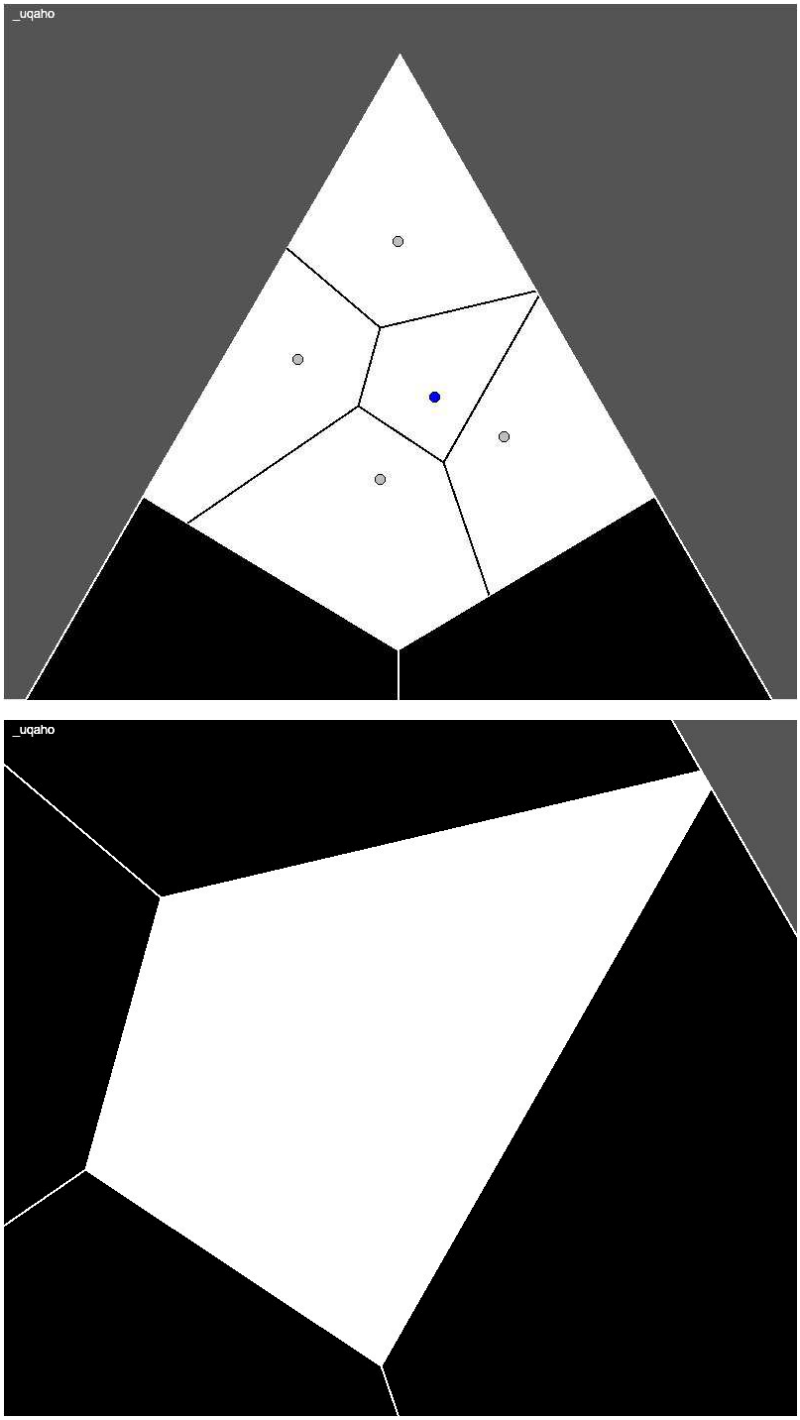

Figure 8.11: Zooming in: the highlighted concept at top is "zoomed in on" at bottom.

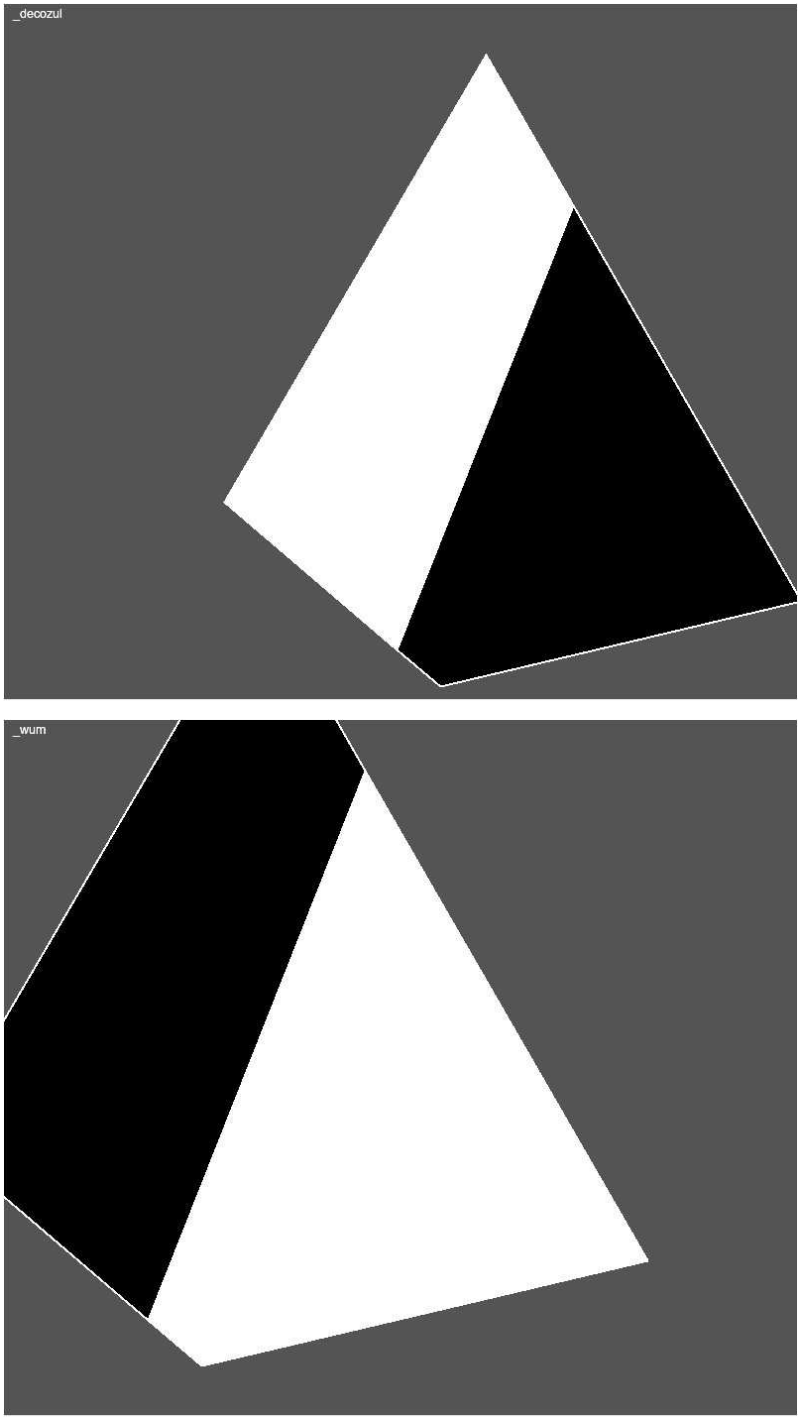

Figure 8.12: Left clicking on a neighbouring black area (unhighlighted) allows you to "scroll" left-right-up-down. Grey areas are "undefined".

8.2.3 Relation to Conceptual Spaces and the Unified Conceptual Space

The program implements many, though by no means all, aspects of the conceptual spaces and unified conceptual space theories described in Chapter Six. Movement is permitted along three of the four axes describing the unified space. Progressive partitioning of that space is facilitated, as are linkages to distal parts of the space.

1. Each such point represents a Voronoi site and the prototype for that concept (Section 6.1.3)[4].
2. Each such re-partitioning is according to the rules for a Voronoi tessellation and each sub-region represents a Voronoi cell. The sub-region represents the range of that concept.
3. Each such move is along either the axis of abstraction (Section 6.2.1.4) or the axis of dynamics (Section 6.2.1.3).
4. The label is a word that gets conventionally attached to the concept (Section 6.2).
5. The links are either components (where applicable), parameters, or contextuals (Section 6.2.2).
6. The zooming moves along the axis of generalization (Section 6.2.1.1).
7. Again, each such move is along either the axis of abstraction or the axis of dynamics.

8.2.4 Limitations and Possible Extensions

8.2.4.1 Theoretical Issues

Movement along the axis of alternatives (adjusting the value of one or more parameters according to some unit measure; see Section 6.2.1.2) is not currently possible. Without this, it is not possible to reconstruct so simple an example as Gärdenfors' (2004) oft-mentioned colour cone. In order to implement this, among other things a distinction needs to be made between the minimal perceptually distinguishable change between parameter values, and the minimal *conceptually* distinguishable change. (See the discussion at the start of Chapter Six.) A lot of thought needs given to how things might best be visualized, while permitting movement by either minimal unit (in portions of the space where such is appropriate), according to any one or more parameters.

There is no proper distinction at present between components, parameters, and contextuals (see Section 6.2.2), other than to make sure they are of the appropriate proto-conceptual type (i.e., objects, action/events, or properties). In particular, no ordering is imposed on components, either on physical objects with respect to physical space or on action/events with respect to temporal ordering. The spatial relation of handle to door should matter, for example, as should the temporal relation of drawing in a rapid breath to sneezing. In addition, no distinction is made between *necessary* components and parameters on the one hand, and *optional* components, parameters, and contextuals on the other.

It is not currently possible to delete a node or to delete a distal link (component, parameter, or contextual) once added (see Section 7.3.2). In consequence, it is not possible to remove partitioning from any area of the unified space. Re-partitioning is supported to a limited extent: one can move points around to force re-partitioning, but while the boundaries of the child nodes are correctly updated, the locations of the central (prototype) points are not. Also, update of boundaries and central (prototype) points for grandchild, etc., nodes does not take place. Ideally, a change at any level of the generalization hierarchy should force a cascade of changes all the way down to the base level.

[4]Compare this list item for item to the list in Section 8.2.2.

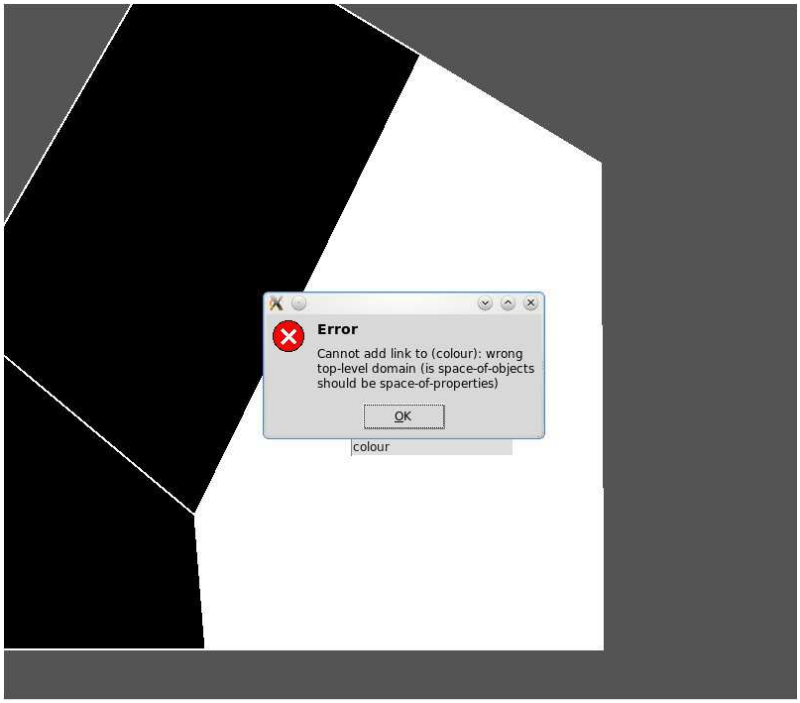

Figure 8.13: Minimal consistency checking: an attempt to make "colour" a component of an object concept is rejected, because "colour" has already been defined as a property concept, and the components of object concepts must themselves be object concepts.

It is not possible for two nodes to have the same name (*polysemy*). It is not possible for a node to have more than one parent (i.e., to be understood relative to more than one domain). A mechanism would need to be provided for choosing between multiple parents when scrolling "upward" along the axis of generalization.

Consistency checking is present but very limited (see Figure 8.13). It would be helpful if the application could, as the original proposal for it suggested, point out logical inconsistencies that are several or more steps removed from each other and which the user may not, as a consequence, have recognized. Turning that around, it could usefully make comparisons between structurally similar concepts. There is currently no mechanism for metaphorical linkages between concepts in different domains. (See the discussion at the start of Chapter Six).

At the moment there is no help system, nor any legend to explain the different mouse options: either what they are or what they mean.

It would be useful if the application provided two, simultaneous views: one focusing on the current concept and its children, the other showing a larger-scale overview with arcs (analogous to the arcs in current mind-mapping and concept-mapping software) depicting the various, and various sorts of, distal links.

8.2.4.2 Technical

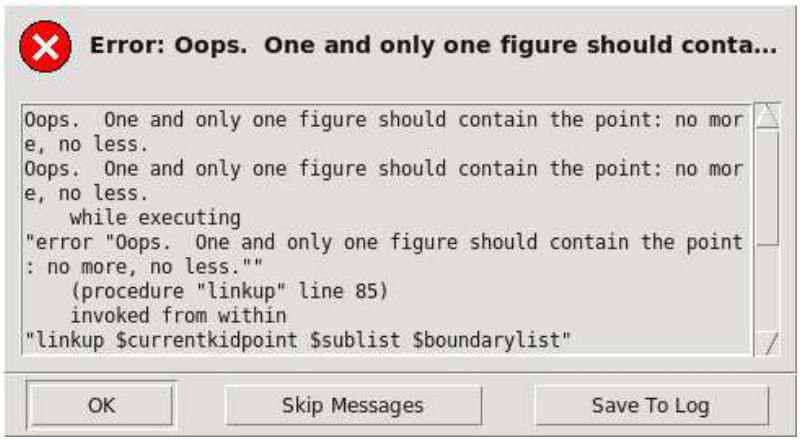

Figure 8.14: Error message: problems correctly locating the boundaries. Here, the program cannot decide which of two possible polygons the node in question belongs to and gives up.

The application is still prone to annoyingly frequent crashes on encountering unexpected situations, although these are gradually being removed. If one forces re-tessellation of a parent concept, the borders of the children are re-calculated, but because their central points are not as yet, this can cause unexpected behaviour or a crash. The algorithm for calculating the borders of child concepts is not quite correct and occasionally makes mistakes (see Figure 8.14). Speed becomes an issue when one has a large number of child concepts, all of whose borders need re-calculated.

All of the nodes are held in a single hash table. This has proven to be a poor choice, and an indexed list would be better. Furthermore, a large (extensively partitioned) space will lead to an unmanageably large hash table or list. It would be much better if portions of the space were written to file or loaded into memory on demand.

It should be possible to put the application on the Web using the Apache Rivet web-server module. So far, however, this has only briefly been explored.

Like all scripting languages, TCL/TK trades speed of prototyping for efficiency. The best way forward would be to re-implement everything from the ground up in TCL/TK then, once a sufficiently stable prototype is achieved, re-implement it in a more efficient language.

8.2.5 Theoretical Significance

8.2.5.1 From Theory to Implementation, Back to Theory

For all of the importance of solid philosophical grounding, I share the distrust of many cognitive scientists, and particularly of AI researchers, toward "armchair" philosophy. In Section 6.1.3, I wrote in general terms of the value of a tight loop from theory to theoretical model to implementation, and back to theory again.

Programming Charley has already resulted in a number of minor changes to the description of the unified conceptual space theory in chapters Six and Seven and in (Parthemore and Morse, 2010), including re-labeling of the proto-concepts, reorganization of much of the material, and re-drawing of several of the figures (7.5, 7.6, and 7.7). It has resulted as well in at least one

major theoretical revision: the description of the unified space in four dimensions instead of three, with the replacement of an *axis of similarity* (from "maximally similar" to "maximally different") by an *axis of alternatives*; the replacement of an *axis of perspective* (from "maximally broad" to "maximally narrow" description) by an *axis of abstraction*; and the addition of an *axis of dynamics*, which previously had been conflated into the *axis of similarity.*

Writing the application has driven home, as well, the difference in unit measures along the different axes. For any two nodes related to each other in a parent/child relationship (axis of generalization*)*, the distinction between parent and child is strictly binary: *either* the one is the parent of the other, *or* it is the child. It cannot be both. This makes for an unambiguous stepping from parent to child to grandchild, or from child to parent to grandparent, and so on. Likewise for any two nodes related to each other solely by a difference in one or more integral dimensions (axis of alternatives), the unit distance will be determined by the minimal perceptually or conceptually distinguishable difference.

In contrast, location of concepts along the other two axes is much more vaguely defined, making correct placement of concepts along these axes difficult. Of course, some concepts are more explicitly and intrinsically dynamic than others, and so one can distinguish object concepts from action/event concepts, and one can take any particular object concept or action/event concept and push it in one direction or the other; but the distinctions are broad, the notion of unit measure nearly irrelevant.

Likewise, some concepts are more clearly toward first-order, others more clearly toward second- and higher-order; of the latter, some concepts are only implicitly second- (or higher-) order while others are explicitly so. One can distinguish property concepts from object or action/event concepts by their being more toward second- and higher-order; and one can take any particular concept, be it object, action/event, or property concept, and push it more toward first-order or more toward second- and higher-order. However, once again, the distinctions are broad. A notion of unit measure suggests itself (zeroth-, first-, second-, and nth-order), but, as I argued already in Section 4.1.1, the boundary between levels is impossible to pin down anywhere. Such reflections suggest that partitioning along these two axes should, at the least, be severely constrained.

8.2.5.2 Toward a More Mature Formalism

Both the theory and the associated KR formalism are, for now, very much in flux. A more mature formalism, resulting from several further iterations of the theory-model-implementation-theory loop, would have a number of practical benefits.

First, to the extent that the formalism continues to prove theoretically interesting and psychologically relevant, it may find application in cognitive science models and AI applications, wherever there is a need to represent knowledge explicitly and motivation to ground that representation on explicitly articulable and solid theoretical foundations. That was, after all, one of the primary motivations for this research project. Recall the quote from Fodor that opened Section 1.3. Every representational approach has consequences, and representing knowledge in terms of conceptual spaces is not like representing knowledge in XML.

Second is the question of embodiment. Back in Chapter One, I said that rather than being seen as unembedded and disembodied, Charley should be understood as offloading embeddedness and

embodiment (and, indeed, its dynamics) onto the user. Perhaps this should be taken as an IOU. That the user provides all of the dynamics in the system should not be taken as a problem – for now. With a more mature formalism and an autonomous robotics platform, one could properly begin to explore issues of salience: not just articulating meaning but *making* meaning; not just drawing a map of the territory in the way that Charley currently makes possible, but creating the territory at the same time as exploring it. This is to say: the current application explores only the *application* side of concept acquisition/application; a robotic platform would allow one to begin to explore the acquisition side (and discover whether, indeed, the same means of representing knowledge can be employed for both, as I have claimed).

8.3 Conclusions

In this chapter I have surveyed the existing field of mind mapping and mind-mapping software and explored its critical limitations. I have presented a mind-mapping application of my own that is strikingly different, both in terms of its superficial visual appearance and operation, and in terms of its theoretical commitments.

The general rule of thumb, introduced in Section 6.1.3, is that theories translate imperfectly to models and models imperfectly to implementations. That said, Charley represents a remarkably good translation of the unified conceptual space theory, in large part, no doubt, because the theory was developed all along with an eye toward how it might be implemented. Adding the ability to move along the fourth axis, the axis of alternatives, would fill in those portions of the original (and underlying) conceptual spaces theory that are missing.

At the same time, the application is very close to providing a serious alternative to existing mind-mapping software, one that is informed by a specific theory of concepts. Existing software, I have argued, is massively under-constrained by any such theory. Too many constraints, of course, can be just as bad as too few. With appropriate constraints, however, the user can be guided in certain directions and away from others, without any need to understand how or why or even that this is happening.

While mind maps have, as Sharples observed, been invested with almost miraculous powers, mind-mapping software to date has failed to live up to the hype, its usage limited to niche user communities. (The software is extensively used, for example, with certain learning disablities, even though there are no reliable statistics to quantify the benefits.) Charley shows how mind mapping can begin to live up to its promise. It invites users to look inside the nodes that existing software treats as, essentially, black boxes: i.e., to understand them as having not just distal connections but internal structure.

Of course, it may be the case that the "maps" produced with Charley will be no more similar than those produced with existing software, even between experts in a domain; but if superficial similarities are lacking (as one might expect, because of the public/private distinction I raised in Chapter Three), nonetheless, underlying structural similarities should allow one to see how these different, individualized spaces come together in a shared conceptual space of the group or the society. This is to say, the maps produced by different users should be more readily understandable and require less explanation from the person who drew them.

Finally, no claim has been made either that the unified conceptual space theory or the software application based on it is complete. An online version of the application that invites experimentation would be a good way to begin to gather data, reveal gaps in the theory, and push theory and application in the direction of formal empirical testing.

Chapter 9

Conclusions and Future Work

I think that the single most important contribution of this research project is the way it sets out the distinction between two contrasting perspectives on concepts and two, corresponding, approaches to understanding them; then demonstrates the way that most contemporary theories of concepts line up on one side of the divide or the other. Are concepts ("mental") representations, or are they (non-representational) abilities? They must be both. The seeming inconsistency between the two perspectives is nothing more than a reflection of the unavoidable limitations of our conceptual horizons. A version of conceptual spaces theory is, I believe, best placed to bridge the divide and bring the two sides as close together as is practical (and, indeed, either possible or desirable) to do. The principal contributions of each chapter are:

Chapter One: A unique treatment of literal versus metaphorical boundaries as they apply to cognition in general and concepts specifically; a novel argument for the importance of metaphysical biases to the discussions in the remainder of the thesis.

Chapter Two: The distinction between concepts as we reflect upon them and concepts as we possess and employ them non-reflectively – part of what I later (Chapter Five) call the *toggling effect*; a novel treatment of representations, and the relationship between concepts and representations, tied to where one locates the observer.

Chapter Three: The distinction between the private and public aspects of concepts, and an argument why that should be understood in terms of contrasting aspects as opposed to different kinds of concepts; a novel argument for including *evolvability* among the core properties of concepts, and against including *introspectibility* and *articulability*, as commonly understood.

Chapter Four: A novel classification of concepts by types of conceptual agents, types of referents, and manner of use; a novel application of Donald's four stages of cognitive-cultural evolution to the evolution of concepts, addressing an oft-neglected issue in theories of concepts.

Chapter Five: A novel elucidation of the relevance of self-reference to theories of concepts; the notion of concepts as *necessary fictions*; a novel argument in favour of the extended mind hypothesis that avoids the risk of cognitive bloat; a novel response to Penrose's argument that human understanding is not bound by Gödel's Incompleteness Theorem.

Chapter Six: The unified conceptual space theory as an extension of conceptual spaces theory, showing how all of an agent's many conceptual spaces come together into a single space of spaces; a clarification of the relationship between conceptual spaces theory and enactive philosophy.

Chapter Seven: A detailed theoretical account of the co-emergence of concepts and experience, as an application of conceptual spaces theory and the unified conceptual space theory; the notion of *concepts as expectations*, increasingly structuring the experience that structures them.

Chapter Eight: A software implementation of many aspects of the unified conceptual space theory and illustration of the benefits of such an application to the further iterative development of the theory; at the same time, a visually and theoretically distinctive approach to mind mapping and non-linear structuring of ideas.

9.1 Looking Back

It might be useful at this point to consider the research proposal I wrote five years ago, as I was embarking on this project[1].

> I see my thesis being a specification of a KR formalism and the creation of a toy environment as a springboard for philosophical discussion. The purpose of such a formalism can be left intentionally ambiguous, at least initially: are we merely representing the conceptual knowledge of agents or are we representing the way that knowledge is represented by the agents themselves? We can simply say: this is a way of representing conceptual knowledge (or some appropriate sub-domain?) with a uniform representation; let's see what the consequences are. Areas for philosophical exploration:
>
> - What is the essential nature of "concept"? Though representation seems inevitable, the view of conceptual knowledge as a set of representations is not. What is the value for suggesting a uniform representation for concepts? Is a truly uniform representation (a) formally describable and (b) even possible?
> - We have our notions of "real world" (or "external world") and "mental world", and in popular discourse we generally treat them as being quite separate and distinct. But medical and psychological research would seem to indicate that there is never any uninterpreted ("unmediated") translation of information from the external world to the mental one. So: what is the nature of the relationship between external world "object" and mental world "model"? Are they necessarily (a) separate and distinct; (b) related, possibly even isomorphic; or (c) simply different levels of the same thing (e.g., "model", "model of the model", "model of the model of the model", and so on).
> - Human language is, of course, another form of representation. We can say that we use it to mediate between our own mental worlds and our own models of the external world on the one hand, and other people's mental worlds and their models of the external world on the other. Linguists at least since Chomsky have postulated a common structure to all human language; would this structure be related in any way to how the mind itself structures knowledge, chosen specifically because of that similarity?

[1] This version has been shortened from the original.

- If we take as starting point (for the sake of argument) that "mind" consists of some set of representations, what (if any) is the usefulness of viewing "consciousness" as just another representation, but one that is unique insofar as being a representation of the system itself (thereby creating what Hofstadter has been fond of calling "tangled loops")?

Looking more closely at the nature of "concept", what understanding am I starting from? Premises:

- Considering concepts as mental representations: "... Every aspect of thinking can be viewed as a high-level description of a system which, on a low level, is governed by simple, even formal, rules" (Hofstadter, 2000, p. 559).
- As a general rule, concepts can be viewed as an "atom" or broken down into smaller concepts. That is to say, any concept has certain requisite components. (My concept of "man" requires a "head" and a "torso"; otherwise, it's not a "man". Other components, such as two arms and two legs, are attached to my concept of "man" but not requisite.)
- At some point, this decomposition has to stop. One can postulate an atomic "concept" from which one will derive all other concepts. This is the "uniform representation" suggested above: the underlying "building block". (Philosophers probably would not want to call such a low-level construct a "concept" at all.)
- Any concept has certain requisite descriptive properties or parameters. (My concept of "man" requires "man" to be "alive"; otherwise the appropriate concept is something different, like "corpse".)
- Any concept exists within a certain context. Any concept that has no context has no meaning.

Most of the themes of the present work are already reflected here: the relationship between concepts and representations, concepts and language, concepts and proto-concepts, concepts and referents, concepts and context; the search for a uniform representation formalism; the questions of conceptual ontologies, conceptual properties, and metaphysical perspectives; the plans for a software-based test application.

If I were to criticize myself, it would be for attempting to do what I claimed in Chapter 5 I could not: provide a too-complete theory that tries too hard to be universal. More practically, I would criticize myself for allowing the philosophy of mind to play too dominant a role in the final product and failing to maintain a balance between largely armchair-based theorizing and hands-on application: one of the very tensions I am claiming to be most important to hold onto (Section 6.1.3). AI researchers will not see the immediate relevance to how they do "knowledge representation" – the thesis offers no neat module that can simply be slotted into their work; cognitive scientists, unless they are already in the grip of Fodor's angst (Section 1.3), probably not see the motivation for making implicit assumptions about concepts explicit. Meanwhile, the debate between concepts-as-(mental)-representations and concepts-as-abilities will go on.

As noted at several points, the present account makes no real effort to define salience. Neither does it address conceptual dynamics in any satisfactory way: as noted in Section 6.2.3, a mature unified conceptual space should be able to capture a rich notion of dynamics internally: i.e., it should be able to capture something about the way it and its world model change.

The argument on metaphysics and boundaries was recently accepted for publication (Parthemore, 2011), much to my satisfaction. Chapters Two through Four present, perhaps, the most polished material, having been re-written any number of times since the first draft of what was to become all three chapters was written almost exactly five years ago; but the argument in Chapter Five, for all that it has gone through several versions itself, still is not so clear as it needs to be. Chapter Six, though it represents published material (Parthemore and Morse, 2010), sets out the framework for the unified conceptual space theory neither to my own nor Peter Gärdenfors' satisfaction; instead, he says "it's getting there". Chapter Eight is much briefer than I would have cared for, and the software program itself is, though visually quite suggestive, still some distance from being put to practical use. That said, it has usefully revealed a number of areas in which the unified conceptual space theory needs to be pressed forward if a truly unified space of all the conceptual spaces is to be achieved.

9.2 Looking Ahead

> People are willing to rank simple objects of different shape and colour on the basis of "similarity". If machines are to reason about structure, this comparison process must be formalized (Aisbett and Gibbon, 1994, p. 143).

Like all theses, this thesis is a work in progress. The notions of *concepts as necessary fictions* (Section 5.2.3), *the toggling effect* (Section 5.4), and *concepts as expectations* (Section 7.3.1) in particular could all use further elucidation, as could the public/private distinction (Section 3.3.3). The "response to Penrose" (Appendix A) could use further polishing and, if it continues to stand the test of argument (and is sufficiently distinct from McCullough's own argument (McCullough, 1995)), submission for publication.

To my mind, the most exciting avenue for further research lies with development of the test application presented in Chapter Eight. Just because theories of concepts are not amenable to direct empirical testing does not mean that they cannot be tested in a variety of extremely productive *indirect* ways. Empirical testing based on "Charley" could include:

- Exploring how natural (or unnatural) naive users find the application, as (indirect) evidence for some isomorphic relation between how it structures information and how their underlying conceptual thought processes do so.
- Exploring how naive users make sense (or fail to make sense) of a map created by someone else for a particular domain: how they "correct" it, etc.
- Exploring how experts judge maps created by novices.
- Comparing how different users (experts and novices) carve up a particular domain.
- More broadly: discovering shortcomings in either conceptual spaces theory or the unified conceptual space theory.
- More ambitiously: automatically extracting conceptual dimensions from a suitable corpus using some version of *latent semantic analysis* (LSA)[2].

[2]For a good introduction to LSA, see (Landauer et al., 1998).

- Most ambitiously: exploring how, *per* the brief remarks at the end of Chapter Eight, the application could be given some degree of autonomy, e.g. by using a robotic platform, and so close the circular causal loop presented in Chapter Seven. Such an autonomous system would need some way of automatically extracting conceptual dimensions according to some measure of salience and some way of automatically generating the metric for the resulting spaces in a way such as Aisbett and Gibbon have suggested. (See the discussion of metrics in the introduction to Chapter Six.)

That said, before such an autonomous platform can even begin to be explored, the theory must be more developed and stabilized. As in so many other application areas, the top-down-driven approach provided by the mind-mapping application must come first, in order to constrain meaningfully the subsequent search space.

As I set out in the introductory chapter, a primary motivation for this thesis was the conviction that theories of concepts are both central to cognitive science research and much if not most of the time left implicit. I would hope my thesis can provide some motivation in cognitive science not so much toward adopting *my* theory of concepts as explicitly acknowledging *some* theory of concepts. Concepts (as the structuring units of [essentially] all of our structured thoughts) are foundational, and foundations matter!

Appendix A

Penrose's Argument, and a Response

Usefully, I have no need to rely so directly on intuitions to show why I believe Penrose's argument, the heart of which is neatly summarized in pages 72-77 of ***Shadows of the Mind*** (1994), to be seriously misleading. That is because Penrose himself retreats from arguments based on "obvious" truths, which he makes earlier in the book, to the seemingly safer ground of logic. (Further, he limits the required logic to no more than an introductory university-level understanding.) I will put Penrose's argument in my own words but be at pains (I trust) not to change it.

A.1 Penrose's Argument Step-by-Step

Penrose's argument is made via a version of the well-known halting problem, which is generally accepted to be algorithmically undecidable. Consider a set of functions $\boldsymbol{C}$ all of whose members take as input an arbitrary natural number n and produce some other natural number x as output, by iterating over the natural numbers starting with zero and returning the first acceptable value of x that they find. Some possible members of $\boldsymbol{C}$ are:

- $C_1(n) = x$ where x is odd and $x > n$.
- $C_2(n) = x$ where x is odd and is twice the value of n.
- $C_3(n) = x$ where x is not the sum of n squares.

It should be clear from grade school mathematics that the first function quickly terminates given any value of n and the second never does, regardless of n. On the other hand, the third one (Penrose's example) terminates only on 0, 1, 2, and 3 (according to the mathematicians), which is probably *not* immediately obvious.

What this means is that certain members of $\boldsymbol{C}$ will terminate regardless of the value of n; certain members will fail to terminate regardless of the value of n; and certain members will terminate only on certain values of n. Let us refer to any specific member of $\boldsymbol{C}$ as C_{q}, since we could, in principle, list all the possible members of $\boldsymbol{C}$ and assign a natural number q as an index to each. It is essential to Penrose's argument that for any value of n, the listing should include every possible function that can take a single value n as input and return x. It is likewise essential that for any particular C_{q}, there is no value of n that it cannot take as input.

Next, consider another function A. A is a two-argument function that takes as input some C_{q} – namely, the qth C_{q} – and some value of n and terminates with an output of, say, 1 for "success" (the output really does not matter), if and only if that particular C_{q} does *not* terminate for that value of n. Otherwise, A does not terminate. It is important that A is sound (it does not make mistakes) and knowably sound (it can be strictly relied upon not to make mistakes).

Penrose suggests that we take A to encapsulate "*all* the procedures available to human mathematicians" (1994, p. 73) for deciding whether or not $C_{\mathrm{q}}(n)$ terminates. Two things are essential here:

1. If $A(q,n)$ stops then $C_{\mathrm{q}}(n)$ does not stop.
2. If $C_{\mathrm{q}}(n)$ stops then $A(q,n)$ does not stop.

Note that the following do *not* hold: i.e., they do *not* follow from the previous two statements.

3. If $C_{\mathrm{q}}(n)$ does not stop, then $A(q,n)$ stops.
4. If $A(q,n)$ does not stop, then $C_{\mathrm{q}}(n)$ stops.

That is to say, the failure of either $A(q,n)$ or $C_{\mathrm{q}}(n)$ to stop says nothing at all about whether the other will stop. This asymmetry is likewise critical, as shall be shown below.

What happens when $q = n$ (as must be a possibility, given that there are no restrictions on the value of either q or n)? In that case, we can re-write (1) as:

5. If $A(n,n)$ stops than $C_{\mathrm{n}}(n)$ does not stop.

Remember, however, that $\boldsymbol{C}$ contains every possible C_{q} that takes some n as input and returns some x. Since $A(n,n)$ is now a function of one argument, it must be on the list. If it is the kth element of the list, then, in that case, n takes the value k and:

6. If $A(k,k)$ stops than $C_{\mathrm{k}}(k)$ does not stop.

At the same time, because $A(k,k)$ is the kth element of the list, $A(k,k) = C_{\mathrm{k}}(k)$. Substituting back into (6), one gets:

7. If $C_{\mathrm{k}}(k)$ stops then $C_{\mathrm{k}}(k)$ does not stop.

If this looks like one half of the "eternal oscillation between two competing and contradictory points of view" I have discussed at several earlier points, it should. The successful termination of $C_{\mathrm{k}}(k)$ yields a contradiction. "From this", writes Penrose, "we must deduce that the computation... $[C_{\mathrm{k}}(k)]$ does *not* in fact stop" (1994, p. 75). Indeed, we can do so without any contradiction.

A.2 The Fly in the Ointment

> In order for Penrose's argument to go through, he needs to make the following assumptions about human mathematical reasoning: -Human mathematical reasoning is sound. That is, every statement that a competent human mathematician considers to be "unassailably true" actually is true. -The fact that human mathematical reasoning is sound is itself considered to be "unassailably true" (McCullough, 1995, p. 3).

What stops Penrose, and the rest of us, from completing the loop and being caught in the "eternal oscillation"? Why does (8) not follow (i.e., how is this situation crucially different from Russell's Paradox or the Epimenides Paradox)?

8. If $C_{\mathrm{k}}(k)$ does not stop then $C_{\mathrm{k}}(k)$ stops.

It is, as Penrose properly concludes, the incompleteness of A: i.e., its inability to produce *every* correct answer. A's completeness would trap us in the very paradox from which Penrose wishes to show we are free:

> Thus, our procedure A is incapable of ascertaining that this particular computation... $[C_{\mathrm{k}}(k)]$ does not stop even though it does not. Moreover, if we *know* that A is sound, then we *know* that... $[C_{\mathrm{k}}(k)]$ does not stop. Thus, we know something that A is unable to ascertain. It follows that A *cannot* encapsulate our understanding (Penrose, 1994, p. 75).

Unfortunately for Penrose, whether that last part follows depends quite critically on what exactly it is that we know. It depends, further, on what we *know* that we know[1]: to know without being aware of that knowing, or only to *think* that we know, will not be enough. There cannot be possibility for error on our part in reaching our conclusion about $C_{\mathrm{k}}(k)$.

I take it as untendentious that A cannot be a member of $\boldsymbol{C}$ – if A is sound. (Remember Grush and Churchland's point that A might be benignly unsound.) The potential controversies arise, as Penrose is aware, with how A is understood. McCullough is concerned, as I am, with the ambiguities here. So, for example, the various C_{q}s can be understood, again untendentiously, as Turing machines; but Penrose considers actual, real-world computers *just to be* universal Turing machines, something to which I have objected.

Earlier Penrose invited us to consider that A encapsulates "*all* the procedures available to human mathematicians". What Penrose wants us to conclude, of course, is that A encapsulates only the computational (or *algorithmically describable*) procedures. Mathematicians (and the rest of us) have access to a different, non-computational, function that includes non-computational as well as computational procedures and is able to decide what A cannot. Call this new function R. If A and R are thought of, for sake of argument, as sets of procedures, then A is contained within R: $A \subset R$.

It would seem to follow, however, that R must *also* be incomplete in order to break ourselves out of the "eternal oscillation": i.e., Penrose's argument trades on the incompleteness not just of A but of R. Why is this? If R could produce every correct answer, and irrespective of diagonalization arguments[2], it would follow that:

9. If $R(q,n)$ stops then $C_{\mathrm{q}}(n)$ does not stop (equivalent of (1); required by soundness).
10. If $C_{\mathrm{q}}(n)$ stops then $R(q,n)$ does not stop (equivalent of (2); required by soundness).
11. **If $\boldsymbol{C_{\mathrm{q}}(n)}$ does not stop, then $\boldsymbol{R(q,n)}$ stops (equivalent of (3)).**
12. **If $\boldsymbol{R}$ does not stop, then $\boldsymbol{C_{\mathrm{q}}(n)}$ stops (equivalent of (4)).**

[1] (Chalmers, 1995) echoes this concern.

[2] This is also what I understand to be Daryl McCullough's conclusion: see (1995, pp. 3-4). By subtly different argument, Chalmers (1995) argues that what I am calling R might be sound but cannot be (as Penrose requires) *knowably* sound.

Such is sufficient, I believe, to construct a Liar's-Paradox-type oscillation. Either R is, itself, a member of $\boldsymbol{C}$ (and hence unsound) or R is (perhaps benignly) unsound.

As Daryl McCullough puts it, "... the Gödel argument doesn't prove that human reasoning must be noncomputable – it only proves that if human reasoning is computable, then it must either be unsound, or it must be inherently impossible for us to know both what our own reasoning powers are and to also know that they are sound" (1995, p. 3). Furthermore, "... Even though it might be the case that nothing false is ever judged to be unassailably true, this fact cannot be an unassailable truth" (1995, p. 5).

A.3 So What *Do* We Know?

The lesson I suggest we take from Gödel is precisely that resorting to a larger system – even a partly non-computational one (whatever we take that to mean and I, for one, am unclear, given Penrose's very broad definition) – does not prevent the replacement of one Gödel sentence with another, any more than the messy, informal system we call human language prevents us from constructing riddles like the Liar's Paradox. This is, precisely, Hofstadter's (1979) conclusion. Furthermore, the existence of paradoxes we can see, and not resolve, implies the possibility of paradoxes we cannot see, and therefore cannot consider resolving.

What is it that Penrose's argument actually allows us to conclude that we know? I suggest it is this: that there exist certain specific values of n for which we do, indeed, know that any function that attempts to calculate whether $C_{\mathrm{n}}(n)$ stops will either be caught in an eternally oscillating loop or derive a contradiction. There is no evidence here, and no argument from Penrose:

> *Either* that we can know for ourselves what those specific values of n are. (Of course we could, and would, know those values if R were complete, but R cannot be complete.)
>
> *Or* that A cannot conclude (as a general principle) that there will exist *some* values of n that will produce undecidable propositions.

After all, what Penrose (and Gödel and Turing and Cantor) have shown is only that there are *certain specific* values of n that will produce undecidable propositions.

In conclusion, there is no evidence from Penrose that R is able to do anything that A cannot. In particular, he has not shown that a Gödel-type sentence cannot be constructed within the larger system: to wit, resorting to a "more complete" system does not, of itself, change anything. Is it still possible that R is more complete (or less incomplete) than A? Of course it is – at least until such time as we have an untendentious definition of what is (and is not) computational or algorithmic. What R cannot be is complete (or rather, complete and consistent). Even then, a similarly structured argument gives us reason to think that a *completely* reliable decision procedure to distinguish what is computational from what is non-computational will not be forthcoming.

Bibliography

Abbott, E. A. (1885). *Flatland: A Romance of Many Dimensions*. Roberts Brothers. Also available online from Google Books (http://books.google.com) and from The Gutenberg Project (http://www.gutenberg.org/ebboks/201).

Adams, F. and Aizawa, K. (2001). The bounds of cognition. *Philosophical Psychology*, 14(1):43–64.

Adams, F. and Aizawa, K. (2008). *The Bounds of Cognition*. John Wiley and Sons.

Aisbett, J. and Gibbon, G. (1994). A tunable distance measure for coloured solid models. *Artificial Intelligence*, 65:143–164.

Aisbett, J. and Gibbon, G. (2001). A general formulation of conceptual spaces as a meso level representation. *Artificial Intelligence*, 133:189–232.

Allen, C. (1999). Animal concepts revisited: The use of self-monitoring as an empirical approach. *Erkenntnis*, 51(1):33–40.

Andersen, H. and Nersessian, N. J. (2000). Nomic concepts, frames, and conceptual change. In *Proceedings of the 1998 Biennial Meetings of the Philosophy of Science Association. Part II: Symposia Papers*, volume 67, supplement of *Philosophy of Science*, pages S224–S241. University of Chicago Press.

Ayer, A. (2001). *Language, Truth and Logic*. Penguin.

Baker, G. and Morris, K. (2002). *Descartes' Dualism*. Routledge.

Ballargeon, R. (1987). Object permanence in 3 1/2- and 4 1/2-month-old infants. *Developmental Psychology*, 23(5):655–664.

Barrett, C. (1962). Concepts and concept formation. *Proceedings of the Aristotelian Society*, 63:127–144.

Barsalou, L. W. (1999). Perceptual simulation in conceptual tasks. *Amsterdam Studies in the Theory and History of Linguistic Science*, Series IV, Current Issues in Linguistic Theory(152):209–228.

Barsalou, L. W., Breazeal, C., and Smith, L. B. (2007). Cognition as coordinated non-cognition. *Cognitive Processing*, 8(2):79–91.

Beck, J. S. (2007). The generality constraint and the structure of thought. http://www.webpages.ttu.edu/jabeck/GCST.pdf. Presentation at MindGrad2006 conference, University of Warwick, UK, and unpublished paper.

Berkeley, G. (1999). *Principles of Human Knowledge and Three Dialogues*. Oxford University Press. A Treatise Concerning the Principles of Human Knowledge was first published in 1710.

Bermudez, J. L. (2007a). Uses and abuses of the distinction between conceptual and nonconceptual content. Invited talk at Concepts: Content and Constitution, 11-12 May, University of Copenhagen, Copenhagen, Denmark.

Bermudez, J. L. (2007b). What is at stake in the debate on nonconceptual content? *Philosophical Perspectives*, 21:55–72.

Bierce, A. (1997). The devil's dictionary. ebook: http://www.gutenberg.org/etext/972. First published 1911. Published electronically by Project Gutenberg.

Borges, J. L. (2007). The library of Babel. In Yates, D. A. and Irby, J. E., editors, *Labyrinths*, pages 51–58. New Directions, New York City.

Brandom, R. (2001). *Articulating Reasons: An Introduction to Inferentialism*. Harvard University Press.

Brentano, F. (1995). *Psychology from an Empirical Standpoint*. Routledge.

Brooks, R. A. (1991a). Intelligence without reason. In *Proceedings, IJCAI-91*. Morgan Kaufmann.

Brooks, R. A. (1991b). Intelligence without representation. *Artificial Intelligence*, 47:139–159.

Brown, H. I. (1994). Circular justifications. In *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, volume 1994, pages 406–414.

Burgener, R. (1957). Price's theory of the concept. *Review of Metaphysics*, 11:143–159.

Carroll, L. (1995). What the tortoise said to Achilles. *Mind*, 104(416):691–693. Original publication 1895 in Mind 4(14): 278-280.

Chalmers, D. (2009). The singularity: A philosophical analysis. http://consc.net/papers/singularity-ml.pdf. Version of talk presented at Singularity Summit, New York City, 3-4 October 2009.

Chalmers, D. J. (1990). Why Fodor and Pylyshyn were wrong: The simplest refutation. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pages 340–347.

Chalmers, D. J. (1995). Mind, machines, and mathematics: A review of Shadows of the Mind by Roger Penrose. *Psyche*, 2:1–10. Available from http://www.theassc.org/files/assc/2331.pdf.

Chalmers, D. J. (1996). Facing up to the hard problem of consciousness. In Hameroff, S. R., Kaszniak, A. W., and Scott, A., editors, *Toward a science of consciousness: the first Tucson discussions and debates*, pages 5–28. MIT Press.

Chella, A., Coradeschi, S., Frixione, M., and Saffioti, A. (2004). Perceptual anchoring via conceptual spaces. In *Proceedings of the AAAI-04 Workshop on Anchoring Symbols to Sensor Data*, pages 40–45, Menlo Park, California. AAAI Press.

Chella, A., Frixione, M., and Gaglio, S. (2000). Understanding dynamic scenes. *Artificial Intelligence*, 123:89–132.

Chella, A., Frixione, M., and Gaglio, S. (2008). A cognitive architecture for robot self-consciousness. *Artificial Intelligence in Medicine*, 44(2):147–154.

Chomsky, N. (2006). *Language and Mind*. Cambridge University Press, third edition. First published 1968.

Chow, T. Y. (1998). The surprise examination or unexpected hanging paradox. *The American Mathematical Monthly*, 105(1):41–51.

Chrisley, R. (1996). *Non-Conceptual Content and Psychological Explanation: Content and Computation*. PhD thesis, University of Oxford.

Chrisley, R. (2009). Interactive empiricism: The philosopher in the machine. In McCarthy, N., editor, *Philosophy of Engineering: Proceedings of a Series of Seminars Held at the Royal Academy of Engineering*, London. Royal Academy of Engineering. In press.

Chrisley, R. and Parthemore, J. (2007a). Robotic specification of the non-conceptual content of visual experience. In *Proceedings of the AAAI Fall Symposium on Consciousness and Artificial Intelligence: Theoretical Foundations and Current Approaches*. AAAI Press.

Chrisley, R. and Parthemore, J. (2007b). Synthetic phenomenology: Exploiting embodiment to specify the non-conceptual content of visual experience. *Journal of Consciousness Studies*, 14(7):44–58.

Christiansen, M. H. and Chater, N. (2001). Connectionist psycholinguistics: The very idea. In *Connectionist Psycholinguistics*, pages 1–18. Greenwood Publishing Group.

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy*, 78(2):67–90.

Churchland, P. S. (1989). *Neurophilosophy: Toward a Unified Science of the Mind-Brain*. MIT Press.

Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press. Also available as ebook through Oxford University Press (http://www.oxfordscholarship.com).

Clark, A. and Chalmers, D. (1998). The extended mind. *Analysis*, 58(1):7–19.

Cupitt, D. (1980). *Taking Leave of God*. SCM Press.

Damasio, A. (2000). *The Feeling of What Happens: Body, Emotion and the Making of Consciousness*. Vintage.

Damasio, A. R., Tranel, D., and Damasio, H. C. (1991). Somatic markers and the guidance of behaviour: Theory and preliminary testing. In Levin, H. S., Eisenberg, H. M., and Benton, A. L., editors, *Frontal Lobe Function and Dysfunction*, pages 217–229. Oxford University Press, New York.

Davidson, D. (1980). *Essays on Actions and Events*, chapter 3: Agency, pages 43–71. Oxford University Press.

Davidson, D. (1987). Rational animals. *Dialectica*, 36(4):317–327.

de Vries, W. (1996). Sellars, animals, and thought. http://www.ditext.com/devries/sellanim.html. Presentation to the Eastern Division of APA, December 1996.

Dennett, D. C. (1969). *Content and Consciousness*. Routledge & K. Paul.

Dennett, D. C. (1991a). *Consciousness Explained*. Little, Brown.

Dennett, D. C. (1991b). Real patterns. *Journal of Philosophy*, 88(1):27–51.

Descartes, R. (1996). *Meditations on First Philosophy*. Cambridge University Press. First published 1641 in Latin.

Dickie, I. (2006). Knowing-which without knowing-that. Presentation to the University of Sussex Philosophy Society June 2006.

Donald, M. (1993). *Origins of the Modern Mind: Three Stages in the Evolution of Culture and Cognition*. Harvard University Press.

Donald, M. (1998a). Hominid enculturation and cultural evolution. In Renshaw, C. and Scarre, C., editors, *Cognition and Material Culture: The Archaeology of Symbolic Storage*, pages 7–17. McDonald Institute for Archaeological Research.

Donald, M. (1998b). Material culture and cognition: Concluding thoughts. In Renshaw, C. and Scarre, C., editors, *Cognition and Material Culture: The Archaeology of Symbolic Storage*. McDonald Institute for Archaeological Research.

Donald, M. (2001a). Memory palaces: The revolutionary function of libraries. *Queen's Quarterly*, 108(4):559–572.

Donald, M. (2001b). *A Mind So Rare: The Evolution of Human Consciousness*. W.W. Norton, London.

Elkin, C. and Greiner, R. (1993). Book review: Building large knowledge-based systems: Representation and inference in the Cyc project. *Artificial Intelligence*, 61(1):41–52.

Ellis, B. (2005). Physical realism. *Ratio*, 18(4):371–384.

Evans, G. (1982). *Varieties of Reference*. Clarendon Press. Edited by John McDowell.

Eyles, A. (1973). Intelligence and differential modes of thinking. *British Journal of Educational Studies*, 21(2):149–155.

Fitch, F. (1964). A Goedelized formulation of the prediction paradox. *American Philosophical Quarterly*, 1:161–164.

Fodor, J. A. (1975). *The Language of Thought*. Crowell.

Fodor, J. A. (1987). Why paramecia don't have mental representations. *Midwest Studies in Philosophy*, 10(1):3–23.

Fodor, J. A. (1998). *Concepts: Where Cognitive Science Went Wrong*. Clarendon Press, Oxford.

Fodor, J. A. (2008). *LOT 2: The Language of Thought Revisited*. Oxford University Press.

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. In Pinker, S. and Mehler, J., editors, *Connections and Symbols*. MIT Press.

Fortune, S. (1987). A sweepline algorithm for Voronoi diagrams. *Algorithmica*, 2:153–174.

Frege, G. (1892). Über sinn und bedeutung. *Zeitschrift für Philosophie und Philosophische Kritik*, C:25–50. Original publication.

Frege, G. (1951). On concept and object. *Mind*, 60(238):168–180. Translated by P.T. Geach and Max Black. First published in 1892 as "Über Begriff und Gegenstand".

Frege, G. (1980). *The Foundations of Arithmetic: A Logico-Mathematical Enquiry into the Concept of Number*. Northwestern University Press, Evanston, Illinois. First published in 1884 as "Die Grundlagen der Arithmetik".

Gallagher, S. (2008). Are minimal representations still representations? *International Journal of Philosophical Studies*, 16(3):351–370.

Gallese, V. and Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*, 22(3-4):455–479.

Gärdenfors, P. (2004). *Conceptual Spaces: The Geometry of Thought*. Bradford Books. First published 2000.

Gärdenfors, P. and Williams, M.-A. (2001). Reasoning about categories in conceptual spaces. In *Proceedings of the Fourteenth International Joint Conference of Artificial Intelligence*, pages 385–392. Morgan Kaufmann.

Geuder, W. and Weisgerber, M. (2002). Verbs in conceptual space. In Katz, G., Reinhard, S., and Reuter, P., editors, *Proceedings of SuB6 (Sinn und Bedeutung)*, pages 69–84. University of Osnabrück.

Gibson, J. J. (1986). *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Associates.

Gödel, K. (1931). Unentscheidbare sätze der principia mathematica und verwandter systeme, i. *Monatshefte für Mathematik und Physik*, 38:173–198. Original publication.

Gödel, K. (1995). Ontological proof. In *Collected Works: Unpublished Essays & Lectures, Volume III*, pages 403–404. Oxford University Press, Oxford.

Goldberg, S. and Pessin, A. (1977). *Gray Matters: An Introduction to the Philosophy of Mind*. M.E. Sharpe, Armonk, New York.

Goldstein, K. and Scheerer, M. (1941). Abstract and concrete behavior: An experimental study with special tests. *Psychological Monographs*, 53(2).

Goodman, N. (1976). *Languages of Art: An Approach to a Theory of Symbols*. Hackett Publishing Company, Cambridge.

Grill-Spector, K. and Kanwisher, N. (2005). Visual recognition: As soon as you know it is there, you know what it is. *Psychological Science*, 16(2):152–160.

Grush, R. and Churchland, P. (1995). Gaps in Penrose's toilings. In Metzinger, T., editor, *Conscious Experience*, pages 185–214. Imprint Academic.

Hall-Haro, C., Price, T., Vance, J., Johnson, S., and Kiorpes, L. (2006). Development of object concepts in pigtailed macaque monkeys. Poster presentation, XVth Biennial International Conference on Infant Studies, Westin Miyako, Kyoto, Japan, June 2006.

Harnad, S. (1990a). Category induction and representation. In Harnad, S., editor, *Categorical Perception: The Groundwork of Cognition*, pages 535–565. Cambridge University Press. First publication 1987.

Harnad, S. (1990b). Introduction: Psychophysical and cognitive aspects of cognitive perception: A critical overview. In Harnad, S., editor, *Categorical Perception: The Groundwork of Cognition*, pages 1–25. Cambridge University Press. First publication 1987.

Harnad, S. (1990c). Preface. In *Categorical Perception: The Groundwork of Cognition*, pages ix–x. Cambridge University Press. First publication 1987.

Harnad, S. (1990d). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(3):335–346.

Harnad, S. (2006). Cohabitation: Computation at 70, cognition at 20. http://cogprints.org/4788/ and http://eprints.ecs.soton.ac.uk/12092/. Presentation at Zenon Pylyshyn Festschrift, University of Guelph, Ontario, Canada, 29 April 2005.

Harnad, S. (2007). From knowing how to knowing that: Acquiring categories by word of mouth. http://eprints.ecs.soton.ac.uk/15690/1/kazimierz.ppt. Presentation at Kaziemierz Naturalized Epistemology Workshop (KNEW), Kaziemierz, Poland, 2 September 2007.

Harvey, I. (1992). Untimed and misrepresented: Connectionism and the computer metaphor (CSRP 245). FTP archive currently offline as of April 2011. University of Sussex (UK) Cognitive Science Research Papers (CSRP) series.

Hawking, S. W. (1988). *A Brief History of Time*. Bantam.

Hayek, F. A. (1999). *The Sensory Order: An Inquiry Into the Foundations of Theoretical Psychology*. University of Chicago Press. First publication 1952.

Heidegger, M. (1978). *Being and Time*. Wiley-Blackwell. First published in 1927 as "Sein und Zeit".

Heinlein, R. A. (1983). "—and he built a crooked house—". In *The Unpleasant Profession of Jonathan Hoag*. Ace Books.

Held, R. and Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. *Journal of Comparative and Physiological Psychology*, 56(5):872–876.

Hemeren, P. (2008). *Mind in Action: Action Representation and the Perception of Biological Motion*. PhD thesis, University of Lund.

Hofstadter, D. (2000). *Gödel, Escher, Bach: An Eternal Golden Braid*. Penguin. Twentieth anniversary edition.

Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books, Inc. Original publication.

Holland, O. (2007). A strongly embodied approach to machine consciousness. *Journal of Consciousness Studies*, 14(7):97–110.

Hume, D. (2003). *A Treatise on Human Nature*. Courier Dover. First published 1739-1740.

Jackendoff, R. (1985). *Semantics and Cognition*. MIT Press.

Jacquette, D. (2005). Grelling's revenge. *Analysis*, 64(3):251–256.

Jaegher, H. D., Paolo, E. D., and Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in Cognitive Science*. In press.

Johnson, S. P., Amso, D., and Slemmer, J. A. (2003). Development of object concepts in infancy: Evidence for early learning in an eye-tracking paradigm. *Proceedings of the National Academy of Sciences*, 100(18):10568–10573.

Jolley, K. D. (2007). *The Concept "Horse" Paradox and Wittgensteinian Conceptual Investigations: A Prolegomenon to Philosophical Investigations*. Ashgate, Aldershot, UK.

Kant, I. (1999). *Critique of Pure Reason (abridged)*. Hackett Publishing Company. First published in 1781 as "Kritik der reinen Vemunft".

Kaye, L. J. (1993). Are most of our concepts innate? *Synthese*, 95(2):198–217.

Keil, F. C. and Wilson, R. A. (2000). The concept concept: The wayward path of cognitive science. *Mind and Language*, 15(2-3):308–318.

Ketland, J. (2005). Jacquette on Grelling's paradox. *Analysis*, 65(3):258–260.

Kirsh, D. (1991). Today the earwig, tomorrow man? *Artificial Intelligence*, 47:161–184.

Kuhn, T. (1970). *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago.

Kuhn, T. (1990). The road since Structure. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, volume 2 of *1990*, pages 3–13.

Landauer, T. K., Folz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Langham, M. (1999). *An Investigation of the Role of Vehicle Conspicuity in the "Looked But Failed to See" Error in Driving*. PhD thesis, University of Sussex, UK.

Laurence, S. and Margolis, E. (1999). Concepts and cognitive science. In Margolis, E. and Laurence, S., editors, *Concepts: Core Readings*, chapter 1, pages 3–81. MIT Press, Cambridge, Massachusetts.

Laurence, S. and Margolis, E. (2002). Concepts. In Stich, S. P. and Warfield, T. A., editors, *The Blackwell Guide to the Philosophy of Mind*, chapter 8, pages 190–213. Blackwell.

LeDoux, J. E. (1996). *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. Weidenfeld and Nicholson.

Lenat, D. (2006). Computers versus common sense. Google Video: http://video.google.com/videoplay?docid=-7704388615049492068. Google Tech Talks.

Lenat, D. and Guha, R. (1989). *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Addison-Wesley Longman Publishing Company.

Locke, J. (2004). An essay concerning humane understanding, volume 1. http://www.gutenberg.org/ebooks/10615. First published 1690. Published electronically by Project Gutenberg.

Machery, E. (2009). *Doing Without Concepts*. Oxford University Press.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology*, 58:25–45.

Massaro, D. W. (1990). Categorical partition: A fuzzy-logical model of categorization behavior. In Harnad, S., editor, *Categorical Perception: The Groundwork of Cognition*, chapter 8, pages 254–283. Cambridge University Press, New York.

Maturana, H. (1978). Cognition. In Hejl, P. M., Köck, W. K., and Roth, G., editors, *Wahrnehmung und Kommunikation*, pages 29–49. Peter Lang, Frankfurt. Available online at http://www.enolagaia.com/M78bCog.html, with the original page numbering retained.

Maturana, H. R. and Varela, F. J. (1992). *The Tree of Knowledge: The Biological Roots of Human Understanding*. Shambhala, London.

McCullough, D. (1995). Can humans escape gödel? a review of Shadows of the Mind. *Psyche*, 2:1–9. Available from http://www.theassc.org/files/assc/2327.pdf.

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

Metzinger, T. (2003). *Being No One: The Self-Model Theory of Subjectivity*. MIT Press.

Miller, G. A. (1994). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 101(2):343–352. Abridged version of original publication from 1956 in the Psychological Review 63(2).

Millikan, R. (1998). A common structure for concepts of individuals, stuffs, and real kinds: More mama, more milk, and more mouse. *Behavioral and Brain Sciences*, 21:55–100.

Millikan, R. (2010). On knowing the meaning; with a coda to Swampman. *Mind*, 119(473):43–81.

Morse, A. and Ziemke, T. (2007). Cognitive robotics, enactive perception, and learning in the real world. In *CogSci 2007 - The 29th Annual Conference of the Cognitive Science Society*, pages 485–490, New York. Erlbaum.

Morse, A. and Ziemke, T. (2010). The somatic sensory hypothesis. Unpublished manuscript.

Newell, A. (1980). Physical symbol systems. *Cognitive Science*, 4(2):135–183.

Newen, A. and Bartels, A. (2007). Animal minds and the possession of concepts. *Philosophical Psychology*, 20(3):283–308.

Noë, A. (2004). *Action in Perception*. MIT Press.

Novak, J. and Canas, A. (2008). The theory underlying concept maps and how to construct them. http://cmap.ihmc.us/Publications/ResearchPapers/TheoryUnderlyingConceptMaps.pdf. Technical report, Florida Institute for Human and Machine Cognition.

Novak, V. (2005). Are fuzzy sets a reasonable tool for modeling vague phenomena? *Fuzzy Sets and Systems*, 156(3):341–348.

Oakeshott, M. (1991). *Rationalism in Politics and Other Essays*. Liberty Press, Indianapolis, Indiana, USA.

Osvath, M. (2009). Spontaneous planning for future stone throwing by a male chimpanzee. *Current Biology*, 19:R190–R191.

Osvath, M. (2010). *Planning Primates: A Search for Episodic Foresight*. PhD thesis, University of Lund, Lund, Sweden.

Oyama, S. (2008). Development without roof, without walls, without floor. Invited talk at the Eighth International Conference on Epigenetic Robotics (EpiRob '08), 30-31 July 2008, University of Sussex, Brighton, UK.

Papineau, D. (2006). Phenomenal and perceptual concepts. In Alter, T. and Walter, S., editors, *Phenomenal Concepts and Phenomenal Knowledge*. Oxford University Press.

Parthemore, J. (1990). Charley: A creative, collaborative writing environment for short story design. Master's thesis, University of Sussex.

Parthemore, J. (2011). Of boundaries and metaphysical starting points: Why the extended mind cannot be so lightly dismissed. *Teorema*, 31. in press.

Parthemore, J. and Morse, A. F. (2010). Representations reclaimed: Accounting for the co-emergence of concepts and experience. *Pragmatics & Cognition*, 18(2):273–312.

Parthemore, J. and Taylor, J. (1992). Charley: A linguistic formalism applied to writing environment design. *Intelligent Tutoring Media*, 3(2/3):85–92.

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

Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford University Press.

Pepperberg, I. (1999). *The Alex Studies*. Harvard University Press.

Perry, J. (1986). Thought without representation. In *Proceedings of the Aristotelian Society*, volume 60, pages 137–151.

Piaget, J. (1954). *The Construction of Reality in the Child*. Basic Books, New York.

Popper, K. (1982). Of clouds and clocks. In Plotkin, H. C., editor, *Learning, Development, and Culture: Essays in Evolutionary Epistemology*, pages 109–120. J. Wiley.

Price, H. H. (1969). *Thinking and Experience*. Hutchinson. First published 1957.

Prinz, J. (2004). *Furnishing the Mind: Concepts and Their Perceptual Basis*. MIT Press. First published 2002.

Prinz, J. (2007). Picture this: Concepts are constituted by percepts. Invited talk at Concepts: Content and Constitution, 11-12 May, University of Copenhagen, Copenhagen, Denmark.

Quine, W. (1951). Main trends in recent philosophy: Two dogmas of empiricism. *The Philosophical Review*, 60(1):20–43.

Quine, W. V. (1969). Natural kinds. In Rescher, N., editor, *Essays in Honor of Carl G. Hempel: A Tribute on the Occasion of His Sixty-Fifth Birthday*, pages 5–23. Springer.

Rakison, D. H. and Lupyan, G. (2008). *Developing Object Concepts in Infancy: An Associative Learning Perspective*. Wiley-Blackwell.

Rosch, E. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605.

Rosch, E. (1999). Principles of categorization. In Margolis, E. and Laurence, S., editors, *Concepts: Core Readings*, chapter 8, pages 189–206. MIT Press.

Rousseau, J. J. (2004). A discourse upon the origin and the foundation of the inequality among mankind. ebook: http://www.gutenberg.org/ebooks/11136. First published 1755. Published electronically by Project Gutenberg.

Ruiz-Primo, M. A. and Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching*, 33(6):pp. 569–600.

Rumelhart, D. (1980). Schemata: The building blocks of cognition. In Spiro, R., Bruce, B., and Brewer, W., editors, *Theoretical Issues in Reading Comprehension*. Erlbaum.

Rupert, R. (2004). Challenges to the hypothesis of extended cognition. *Journal of Philosophy*, 101:389–428.

Rupert, R. (2009a). *Cognitive Systems and the Extended Mind*. Oxford University Press.

Rupert, R. (2009b). Critical study of Andy Clark's Supersizing the Mind. *Journal of Mind and Behavior*, 30:313–330.

Russell, B. (1905). On denoting. *Mind*, 14(56):479–493.

Russell, B. (1908). Mathematical logic as based on the theory of types. *American Journal of Mathematics*, 30(3):222–262.

Russell, B. (1923). Vagueness. *Australasian Journal of Philosophy*, 1(2):84–92.

Ryle, G. (1949). *The Concept of Mind*. Penguin.

Schyns, P. G., Goldstone, R. L., and Thibaut, J.-P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences*, 21:1–54.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–458.

Searle, J. (2002). Why I am not a property dualist. *Journal of Consciousness Studies*, 9(12):57–64.

Searle, J. R. (1999). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press.

Sellars, W. (1956). Empiricism and the philosophy of mind. In Feigl, H. and Scriven, M., editors, *Minnesota Studies in the Philosophy of Science, Volume I: The Foundations of Science and the Concepts of Psychology and Psychoanalysis*, volume I, pages 253–329. University of Minnesota Press. Available online from http://www.ditext.com/sellars/epm.html.

Sharples, M. (1999). *How We Write: An Account of Writing as Creative Design*. Routledge.

Shusterman, R. (2008). *Body Consciousness: A Philosophy of Mindfulness and Somaesthetics*. Cambridge University Press.

Simons, D. J. and Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, 28:1059–1074.

Sloman, A. (1985). Strong strong and weak strong AI. *AISB Quarterly*.

Sonesson, G. (2009). Semiosis beyond signs: On a two or three missing links on the way to human beings. http://project.sol.lu.se/fileadmin/user_upload/project/ccs/Semiosis_beyond_signs.pdf.

Stachel, J. (1982). Comments on "some logical problems suggested by empirical theories" by Professor Dalla Chiara. In Cohen, R. and Wartofsky, M. W., editors, *Language, Logic and Method: Papers Derived from the Boston Colloquium in the Philosophy of Science 1973-1980*, pages 91–102. Springer.

Steiner, P. and Stewart, J. (2009). From autonomy to heteronomy (and back): The enaction of social life. *Phenomenology and the Cognitive Sciences*, 8:527–550.

Stewart, J. (1995). Cognition = life: Implications for higher-level cognition. *Behavioural Processes*, 35(1-3):311–326.

Stich, S. (2006). Philosophy, intuition, and culture: An overview of a research program. Presentation to the COGS Seminar Series, University of Sussex, UK, 30 May 2006.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology and the Sciences of Mind*. Harvard University Press.

Thompson, E. and Stapleton, M. (2009). Making sense of sense-making: Reflections on enactive and extended mind theories. *Topoi*, 28(1):23–30.

Torey, Z. (2009). *The Crucible of Consciousness: An Integrated Theory of Mind and Brain*. MIT Press.

Travis, C. (1994). On constraints of generality. *Proceedings of the Aristotelian Society, New Series*, 94:165–188. Published by Blackwell Publishing on behalf of The Aristotelian Society.

Trumbo, D. (2007). *Johnny Got His Gun*. Citadel. First published 1939.

Tulving, E. (1972). Episodic and semantic memory. In Tulving, E. and Donaldson, W., editors, *Organization of Memory*, pages 381–402. Academic Press, New York.

van Gelder, T. (1990). Compositionality: A connectionist variation on a classical theme. *Cognitive Science*, 14(3):355–384.

van Gelder, T. and Port, R. F. (1996). It's about time: An overview of the dynamical approach to cognition. In *Mind as Motion: Explorations in the Dynamics of Cognition*, chapter 1, pages 1–43. MIT Press.

Varela, F. J. and Shear, J. (1999). *The View From Within: First-Person Approaches to the Study of Consciousness*. Imprint Academic.

Varela, F. J., Thompson, E., and Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press.

Watts, A. (1957). *The Way of Zen*. New American Library.

Weinberg, J. M., Nichols, S., and Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics*, 29(1-2):429–460.

Wheeler, M. (2005). *Reconstructing the Cognitive World*. MIT Press.

Whitby, B. (2003). The myth of AI failure (CSRP 568). FTP archive currently offline as of April 2011. University of Sussex (UK) Cognitive Science Research Papers (CSRP) series.

Williamson, T. (1992). Inexact knowledge. *Mind*, 101(402):217–242.

Williamson, T. (1996). *Vagueness*. Routledge (Taylor & Francis Books, Ltd.), London.

Wimmer, H. and Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13:103–128.

Winograd, T. and Flores, C. (1986). *Understanding Computers and Cognition: A New Foundation for Design*. Intellect.

Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus*. Harcourt, Brace & Company. First published in 1921 as "Logisch-Philosophische Abhandlung".

Wittgenstein, L. (2001). *Philosophical Investigations*. Blackwell. Fiftieth anniversary edition. First published (posthumously) 1953.

Woodfield, A. (1994). Do your concepts develop? In Hookway, C. and Peterson, D., editors, *Philosophy and Cognitive Science*, pages 41–67. Cambridge University Press.

Zachar, P. (2000). *Psychological Concepts and Biological Psychiatry: A Philosophical Analysis*. John Benjamins Publishing Company.

Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8(3):338–353.

Ziemke, T. (2007). What's life got to do with it? In Chella, A. and Manzotti, R., editors, *Artificial Consciousness*, chapter 3, pages 48–66. Imprint Academic.

Zlatev, J. (2009). The semiotic hierarchy: Life, consciousness, signs and language. *Cognitive Semiotics*, 2009(4):169–200.