The Metaphysics of Mental Representation

RICHARD DE-BLACQUIERE-CLARKSON

Abstract

The representational theory of mind (RTM) explains the phenomenon of intentionality in terms of the existence and nature of mental representations. Despite the typical characterisation of mental representations in terms of their semantics, RTM is best understood as a metaphysical – more specifically formal ontological – theory whose primary defining feature is stipulating the existence of a class of mental particulars called representations. In this regard it is false, since mental representations do not exist.

My argument is primarily methodological. Using an extended analysis of mereology and its variants as paradigmatic examples of a formal ontological theory, I argue for a 'synthetic' approach to ontology which seeks to form a sound descriptive characterisation of the relevant phenomena from empirical data, to which philosophical analysis is applied to produce a rigorous theory. The value and necessity of this method is proved by example in our discussion of mereology which is shown to be defensible given certain assumptions, in particular perdurantism, but still inadequate as an account of parthood without considerable supplementation. We also see that there are viable alternatives which adopt a more synthetic approach and do not require the same assumptions.

Having effectively demonstrated the value of a synthetic approach in ontology I critically examine the methodology employed by RTM and find it severely lacking. In the guise of 'commonsense psychology' RTM cavalierly imposes a theoretical framework without regard to empirical data, and this results in a severe distortion of the phenomenon of intentionality it purports to explain. RTM is methodologically unsound, and so its commitment to the existence of mental representations is utterly undermined. Furthermore the most attractive aspect of RTM – its semantics – can be separated from any commitment to mental representations existing. Even RTM's strongest advocates lack motivation to believe that mental representations exist.

The Metaphysics of Mental Representation

Richard de Blacquière-Clarkson

PhD Thesis
University of Durham
Department of Philosophy
2010

Contents

Preface

This thesis began life several years ago with a vague sense of unease at a particular view of the mind, and no clear idea of what to do about it. It should not be too surprising, then, that a significant proportion of it is concerned with methodology. The overall aim, however, is also substantive. I aim to demonstrate that there are no such things as mental representations, and so a particular school of thought which uses them to explain how we think about things is mistaken. The study of mental representation has hitherto been almost exclusively concerned with their semantics – how these putative entities acquire meaning – so to the best of my knowledge the metaphysical (and, more specifically, ontological) approach I take to evaluating mental representation is both novel and unique. Hopefully this is a good thing.

Chapter one characterises and explores the Representational Theory of Mind, using Millikan and Fodor's varieties as exemplars of the general type. Chapter two diagnoses the theory's ontological commitment to mental representations, and justifies the necessity of approaching it as an ontological theory. Chapter three begins our methodological discussion, using parthood as an extended example of how ontology should be carried out, while chapter four continues this discussion and draws out two methodological morals. Chapter five applies these to the Representational Theory of Mind, using them to argue that the theory should be rejected. Chapter six salvages Millikan and Fodor's semantics from the rubble, in doing so demonstrating there never was any good reason to suppose mental representations exist.

I will make a small number of assumptions throughout this thesis, mostly explicit. There are two implicit assumptions which run throughout. First, that some form of metaphysical realism is true; at the very least, the existence of some thing is ontologically independent of thought and the possibility of thought. Second, that it is generally desirable to be permissive rather than restrictive in one's ontology, in what one takes to be the furniture of the world. It is easier to prune some redundant branches from a full theoretical framework than it is to

scaffold one which is overly austere. This thesis is just such an exercise in ontological pruning, to demonstrate that mental representations do not exist.

For valuable feedback and assistance on some or all of this thesis, or arguments contained within it, I would like to thank first and foremost my supervisor Jonathan Lowe; also John Hawthorne, Robin Hendry, Rognvaldur Ingthorsson, Matthew Ratcliffe, Jonathan Tallant and audiences in Bergen, Durham, Leeds, Manchester, Rome and Sussex. Thanks also to the Royal Institute of Philosophy who supported part of this work by kindly offering me a Jacobsen Fellowship for 2006-7, and to my wife Jo for her considerable patience and support.

1. What is Mental Representation?

The term is ubiquitous in contemporary discussions of the mind, not only in philosophy but in cognitive psychology and elsewhere. Unsurprisingly, the meaning of 'mental representation' varies significantly both across and within different disciplines and debates, and equivocation between different meanings is not uncommon. We can distinguish weak and strong readings of 'mental representation' as follows:

Weak: a lawyer represents his/her client in court, an MP his/her constituents in the House of Commons, a portrait represents a person or animal, and a road sign represents some geographical feature or other piece of information. In each case, the representation or representative picks out relevant features of the world at large and makes use of them. By analogy we can say that at least some thoughts represent whatever is being thought about, meaning they select and make use of relevant features whilst omitting others. This weak reading of 'mental representation' is intuitively appealing, as thought does appear to be directed at particular objects and in general does not encompass every aspect of what is thought about. The weak reading makes no claim about the nature of *how* thoughts represent objects, and is quite innocuous.

Strong: fully explaining this reading will be the main topic of this chapter, but we can give a summary here. In many or most cases, our ability to think veridical thoughts consists in the existence in our minds of a class of entities called representations. These mental particulars are characterised primarily by two features: 1. Representations bridge the gap between thinkers and whatever is thought about (i.e. between mind and world), and so the semantic properties of a thought (including its truth value) depend upon the existence in the thinker's mind of a suitable representation; 2. Suitable representations are those whose internal structure corresponds appropriately to that of what is being thought about. On the strong reading, mental representations function as surrogates for relevant features of what is thought about: being mental particulars they can enter into characteristically mental relations such as reasoning and inference, whilst their structural correspondence to the

'external' objects of thought preserves the requirement that our thoughts match up to the world outside our own heads.

The strong reading is controversial, albeit widely held. The overall argument of this thesis is that mental representations under the strong reading do not exist, i.e. there is no class of mental entities which fulfil the roles described above and further explored below. We will not be concerned with the weak reading in any way; the aim is to challenge an ontological commitment made by a particular school of thought, not to legislate on the language people may use in describing the mind.

1.1 The Problem of Intentionality

The view that mental representations exist (on the strong reading) is primarily motivated as a solution to the problem of intentionality. The problem is explaining how thoughts can seemingly be directed towards, or be about, specific objects. Rather than thinking 'I want food', we can also desire a specific type of food, or a specific item. Likewise our beliefs and other thoughts can be equally specific, satisfied only by a unique object, individual, event or whatever else. In the philosophy of mind in general, not only theories of mental representation, intentionality is widely (if not quite universally) seen as 'the mark of the mental', that which characterises what it is to be a thought.

Representational theories of mind (RTMs) admit considerable variation, but all aim to explain intentionality by appealing to a structural isomorphism between each mental representation and the object being thought about. The particular structure of any representation will be isomorphic to a single object (e.g. 'that sandwich over there'), or multiple objects (e.g. 'a ham sandwich', 'some food'), which would satisfy the thought. In this way RTMs explain how thoughts can be 'about' particular objects.

Two challenges immediately arise for advocates of RTM: how do mental representations acquire the appropriate structures for this to happen, and how can representations explain the intentionality of thoughts which are either mistaken or about non-existent objects? We will return to both presently.

Using representation as a means of solving the problem of intentionality is a widely accepted strategy, so much so that the two concepts are often taken to be the same. For example:

> To say that all mental states exhibit intentionality is to say that all mental states are representational.
>
> Crane 1995, p.37

> [T]he *problem of representation*: how is it possible for one item to represent another? We might equally call it the … problem of *intentionality …* how is it possible for one item to be about another?'
>
> Rowlands 2006, p.1

> And these mental states very often have intentional content: they serve to represent the world.
>
> Chalmers 2004, p.1 of online version

> The philosophy of *intentionality* asks questions such as: in virtue of what does a sentence, picture, or mental state represent that the world is a certain way?
>
> Williams 2007, p.2 of online version

> It appears increasingly that the main joint business of the philosophy of language and the philosophy of mind is the problem of representation itself: the metaphysical question of the place of meaning in the world order. How can anything manage to be *about* anything[?]
>
> Fodor 1987, p.xi

> 'You can't reduce intentionality to "selection for" because *selection for doesn't involve representation*.'
>
> Fodor 1998b, p.185

This identification is a mistake as intentionality and representation are quite clearly distinct; after all the main purpose of mental representation on the strong reading is to give an *explanation* of intentionality, and a straightforward identification would render this

vacuous. The mistake has been explicitly recognized by Cummins (1981, p.14), who attributes the 'close connection' between the concepts primarily to Fodor. The quotations above notwithstanding, Fodor himself does show some awareness of the distinction when he observes his own theory 'doesn't, of course, *solve* the problem of intentionality; it merely replaces it with the *unsolved* problem of representation' (1998b, p.184-5).

It is sometimes observed that the problem of intentionality may be a pseudo-problem, on the grounds that its historical basis is misunderstood. The argument is that as a misinterpretation of the historical problem, the contemporary problem of intentionality is illegitimate. If correct, this would render RTMs redundant as they would be solving a problem which does not exist.

The problem of how thoughts can be about objects, particularly non-existent objects, is widely attributed to comments made by Franz Brentano in his *Psychology from an Empirical Standpoint*:

> Every mental phenomenon is characterized by what the Scholastics of the Middle Ages called the intentional (or mental) inexistence of an object, and what we might call, though not wholly unambiguously, reference to a content, direction toward an object (which is not to be understood here as meaning a thing), or immanent objectivity. Every mental phenomenon includes something as object within itself...
>
> This intentional inexistence is characteristic exclusively of mental phenomena. No physical phenomenon exhibits anything like it. We can, therefore, define mental phenomena by saying that they are those phenomena which contain an object intentionally within themselves.
>
> Brentano 1995, p. 88-9

This characterisation is ambiguous in a number of respects. In the literature on mental representation it is almost universally interpreted as consisting of two claims. First that intentionality is paradigmatically the ability to think about what does not exist (see e.g. Millikan 1995, p.187), and in all cases it operates in virtue of a mental surrogate, or 'intentional object' (see e.g. Segal 2005, pp.283-4) which makes thought about things

possible. Second that the mental is irreducible to the physical. This latter interpretation is largely due to Chisholm (1957) but also Quine, who sums up the view succinctly:

> One may accept the Brentano thesis either as showing the indispensability of intentional idioms . . . or the emptiness of the science of intention. My attitude, unlike Brentano's, is the second.
>
> 1960, p. 221

That both interpretations' accuracy to Brentano's intentions is dubious has been made clear by Crane (2006), Moran (1996), Simons (1995) and Smith (1994). Brentano was – to use Simons' helpful terminology – a methodological phenomenalist, meaning that he was not concerned with the ultimate ontological status of the objects of intentional thought. Although the analogy is not perfect, we can say that Brentano is only concerned with phenomena in something like Kant's sense, and has no interest in the noumena. Brentano's methodology entirely precluded the questions of whether the objects of thought exist or not, and whether intentionality is irreducibly mental.

So, for example, when Fodor talks of 'removing Brentano's chestnuts from the fire' (1990a, p.70), meaning giving a physical explanation of intentionality, a closer analogy would be to say that Brentano does not have any chestnuts, and is playing cards in another room. The actual existence of any object being thought about is irrelevant to Brentano's concerns.

Does this render intentionality, as understood in contemporary debates, a pseudo-problem? Not at all, as the concerns raised are legitimate: how can thoughts be directed towards objects, particularly those which do not exist, and does this require a surrogate object; also can an explanation of this phenomenon be given in purely physical terms? Although mid-to-late Twentieth Century innovations, together these questions form the contemporary problem of intentionality, which representational theories of mind seek to address.

The contemporary problem of intentionality is legitimate, and while its historical basis is highly valuable in understanding the problem, the problem can itself be considered in isolation. To disregard the contemporary problem because of its origins would be to commit

a clear genetic fallacy. So we will return to how RTMs seek to solve the problem, as an analysis of how thinking works.

1.1.1 Commonsense Psychology and Mental Content

As an analysis of how thinking works, RTMs naturally incorporate a characterisation of what ordinary or everyday thinking involves. This characterisation is called 'commonsense' or 'folk' psychology, and takes ordinary thought to paradigmatically involve prediction and explanation in terms of propositional attitudes:

> Human beings are social creatures. And they are reflective creatures. As such they continually engage in a host of cognitive practices that help them get along in their social world. In particular, they attempt to understand, explain and predict their own and others' psychological states and overt behaviour; and they do so by making use of an array of ordinary psychological notions concerning various internal mental states, both occurrent and dispositional. Let us then consider folk psychology to consist, *at a minimum,* of (a) a set of attributive, explanatory and predictive practices, and (b) a set of notions or concepts used in those practices.
>
> Von Eckhardt, p.300

> It has come to be a standard assumption in philosophy and psychology that normal adult human beings have a rich conceptual repertoire which they deploy to explain, predict and describe the actions of one another and, perhaps, members of closely related species also. As is usual, we shall speak of this rich conceptual repertoire as 'folk psychology'.
>
> Davies & Stone 1995a, p.2

These practices are classically conceived in terms of beliefs and desires (see e.g. Churchland 1988, p.59; Dennett 1987, p.47; Fodor 1987, p.7), but it is generally accepted that while these form the core, commonsense psychology incorporates ascriptions of many other propositional attitudes such as hopes, fears, intentions (see e.g. Gopnik and Wellman, p.240; Heal, p.45) and possibly emotions (see Baker 2001, p.3; Gordon 1995b, p.120; Harris, pp.207-8). [1]

[1] Different conceptions of folk psychology can easily include or exclude various propositional attitudes, and likewise need not be restricted to prediction, explanation and attribution/description as their uses. For

Propositional attitudes admit of two components, a content and an attitude taken towards that content. A belief is always a belief *that* something is the case, a desire is always *that* something should happen. The concept of content does in fact come from Brentano's theory of intentionality, although his use (quoted above) is quite unclear. His student Twardowski clarified it by distinguishing intentional contents from objects; contents are abstract mental entities, whereas objects are typically concrete physical objects.[2] We can also usefully distinguish intentional targets from both contents and objects. Imagine seeing a figure in the distance wearing a hat and mistakenly identifying him as Tom, although it is actually John. When thinking 'Tom has a fine hat on', the content of this belief is that Tom has a fine hat on, its object is John, while its intended target is Tom. Content and target cannot be identified as while they will coincide in cases of veridical thought, it may well be the case that Tom is wearing no hat at all and the belief is false. Even if Tom is wearing a hat and so the belief is true, we still need to distinguish target from object to support the epistemological point that the belief is almost certainly not justified in this case. Targets, like objects, will also be concrete objects in most cases, while contents are not. As a note of caution, contents are also not necessarily to be identified with propositions, on the grounds that propositions may or may not be physical.

While some versions of RTM seek to also encompass cases of non-propositional thought, such as that of animals and young infants and possibly some perceptual content, propositional attitudes remain the paradigmatic cases of human thought which such theories seek to explain:

---

example, Churchland (1989, p.231) has suggested that folk psychology involves not only prediction and explanation, but also manipulation and control. Morton (1996) argues that it should be seen as a passively anticipatory rather than actively predictive device. Such variations nevertheless preserve the defining feature, that folk psychology consists of the ascription of propositional attitudes to individuals to enable interpersonal understanding. A complete account of ordinary thought might be expected to incorporate more domains than just psychology, but this point has been almost entirely overlooked in the literature. While there have been attempts to provide a systematic account of lay reasoning in other areas, such as Hayes' 'naive physics' (1978, 1983) to which lip service is sometimes paid, I know of no serious attempt to integrate such work with folk psychology. I exclude Shanahan's attempt as its concerns are technical rather than philosophical.

[2] *A priori* reasoning provides a plausible exception.

> RTM says that there is no believing-that-P episode without a corresponding tokening-of-a-mental-representation episode, and it contemplates no locus of original intentionality except the contents of mental representations.
>
> Fodor 2003, p.8

> A theory of mental representation is supposed to describe the concept or knowledge structure underlying people's ordinary judgements about the contents of beliefs, desires and other intentional states.
>
> Stich 1994, p.250

> Intentional representations always come with propositional attitudes attached.
>
> Millikan 2002, p.6 of online version

Although widely supported in contemporary philosophy of mind, often to the point of seeming beyond reasonable doubt, folk psychology is a substantive philosophical position which has received some strong criticism in recent years. We will examine whether it is a plausible account of ordinary thought in chapter five, finding that there are strong arguments to think it is not.

1.1.2 Mental Location: Content and Vehicle Externalism

A further contemporary distinction regarding mental content, which we will make further use of in chapter six, is that of content versus vehicle. Key to understanding the difference is the issue of location. Externalists take at least some mental content to be individuated according to objects located outside of a relevant boundary – typically the brain or central nervous system, but sometimes the body – with the classic and most influential arguments for this position being due to Putnam (1975) and Burge (1977).[3] Internalists take the converse view that mental contents are individuated according to objects which lie within the relevant boundary. There are numerous good and recent surveys of the large and relatively complex literature – for example Brown, Lau & Deutsch, Pessin & Goldberg, and

[3] This is not to deny that other motivations for content externalism exist, such as those proposed by Evans (1982) and McDowell (1984). We should of course also note that Putnam's arguments in 'The Meaning of Meaning' were intended to demonstrate that linguistic, rather than mental content 'ain't in the head'.

Rowlands (2003) – so there is no need to rehearse well-known arguments and positions in detail.

Briefly, then, we can observe some salient points. The first is that there need not be such a clear dichotomy between externalism and internalism regarding mental content as the quick description above might imply. While some philosophers such as Burge (1977) and Davidson (1987) – who Segal (2000) refers to as 'extreme' externalists – have argued that all or almost all mental content is external, and others the no less extreme position that it is entirely internal, perhaps the majority in recent years endorse both internal and external content. To do so is, in the standard terminology, to endorse 'wide' or 'broad' content, rather than a purely 'narrow', i.e. internally individuated, variety. This is not, however, to say that all contents are made equally. For example, McGinn (1989) distinguishes between weak and strong externalism, with only strongly external content bearing causal powers, and argues in favour of the weak version as 'clearly and uncontroversially true' (*ibid.* p.36) whereas strong externalism is in most cases false. Thus in general it is narrow content which causally determines our behaviour: 'the self is in the head' (*ibid.* p.46). Similarly, Dennett (1982), Loar and Fodor (1987, ch.2) each argue in favour of a conception of narrow content which alone is sufficient to underwrite psychological ascription, with wide content being formed by supplementing the narrow with environmental factors. However, in essentially the converse position to those just mentioned Stalnaker takes narrow content *per se* to be unnecessary in accounting for both the role of content in psychological explanation and privileged first-person access to one's own mental states, while agreeing that:

> something like Loar's conception of narrow content will help to describe and explain the ways in which our uses of content to characterise the states of mind of ourselves and others are context dependent.
>
> Stalnaker 1990, p.145

Indeed, Fodor (1994) rescinded his earlier view in favour of wide content. Perhaps a broad consensus has emerged in favour of some variety of externalism, which does have the advantage of a lesser 'burden of proof' than a thoroughgoing internalism – sufficiently weakly construed, externalism only requires a single counterexample to internalist

explanation, and in any remotely plausible variety can admit both internal and external factors (see Hurley 2010, fn.2).

More recently, externalism regarding the individuation of mental content has been contrasted with externalism regarding the processes which make thought possible, a position known variously as 'vehicle' (Hurley 1998), 'active' (Clark and Chalmers), or 'locational' (Wilson 2004) externalism. Vehicle externalism challenges the traditional view that the processes which make thought possible are located entirely inside the thinker's body, most likely entirely inside his brain or central nervous system. Instead, it argues that mental processes are in many or most cases at least partly instantiated in things which are outside the body. The seminal source for this view is Clark and Chalmers' paper *The Extended Mind*, where they use a series of thought experiments to support this view. We will consider one of their examples to illustrate the position, whilst remembering the literature on the topic is large and contains many other arguments and a broad range of related views. The example we will consider is that of Otto:

> Otto suffers from Alzheimer's disease, and like many Alzheimer's patients, he relies on information in the environment to help structure his life. Otto carries a notebook around with him everywhere he goes. When he learns new information, he writes it down. When he needs some old information, he looks it up. For Otto, his notebook plays the role usually played by a biological memory.
>
> Clark & Chalmers, quote taken from online version

The argument is that Otto's notebook fulfils exactly the same function as at least part of a person's biological memory in normal cases, on the grounds that it enables the same beliefs and other propositional attitudes, likewise the same behaviour. The processes, or vehicles of thought, which allow Otto to visit a museum (this is Clark & Chalmers' example) are functionally equivalent to those which allow non-Alzheimer's sufferers to do the same. In Otto's case the mental process requires the existence and use of his notebook. Therefore the cognitive apparatus which makes thought possible is not limited to the inside of our bodies or heads.

There are at least two ways to interpret this example (and *mutatis mutandis* the numerous others which have been proposed by Clark & Chalmers and elsewhere). First, it may be regarded as what Dennett has called an intuition pump, a way of drawing out philosophers' intuitions on the topic of where the vehicles of thought are located. On this interpretation the example, and other like it, may draw out the intuition that vehicle externalism applies in many or most cases, that it is intuitively false, or anywhere in between. Analogously with McGinn's weak content externalism, it may be conceded that the presence of external objects is required in some cases, but these objects lack causal powers. On this interpretation thought experiments in favour of vehicle externalism are highly inconclusive.

A second stronger interpretation yields more definite results. Otto's notebook, and other comparable examples, are not thought experiments at all but are demonstrations of the existential dependency relations which hold between patterns of thoughts or reasoning (whether folk psychological or otherwise), and the physical processes which make them possible. I am not assuming physicalism to be true here – however it is abundantly clear from neurological evidence regarding brain damage that certain physical processes must exist and operate normally for cognition to be possible. That much is uncontroversial. The issues here are whether there is a principled metaphysical boundary which delimits what can constitute part of such a process, and whether this boundary lies at the individual's body or brain. Vehicle externalism as proposed by Clark and Chalmers addresses the second part of this issue. Without Otto's notebook, he is unable to follow a specific pattern of thought. With the notebook, Otto is able. This demonstrates that Otto's thought process is existentially dependent upon the notebook as well as Otto's own biological processes, and so neither the body nor brain constitutes a principled metaphysical boundary delimiting what can be part of a vehicle of thought.

Another way of phrasing the point is this: thoughts are only generically existentially dependent upon the physical processes, or vehicles, which make their existence possible. Any other process which fulfils the same function will do just as well. [4] As long as all of the

[4] Generic existential dependence is contrasted with specific cases, where only a single specified object will satisfy the relation. For a highly detailed account of existential dependency, including the generic/specific distinction see Correia; for an overview of the topic see Lowe 2010.

parts of such processes can interact suitably to fulfil their function, their spatial location is irrelevant. As Rowlands puts it, 'there may well exist vehicles of representation inside the skin of representing subjects. But vehicles of representation do not, in general, stop at the skin' (2006, p.24).

Further examples are required to generalise beyond factual recall memory, though there is no shortage in the literature – for the most recent see the papers collected in Menary. I do not take it that our discussion here proves vehicle externalism to be conclusively true or in any way beyond reasonable doubt. Even the brief consideration given above, however, does demonstrate that it is highly plausible on its own merits. If false, it is not clearly or obviously so. We make use of this point when we return to vehicle externalism in chapter six.

The implications of vehicle externalism are also far-reaching, including the possibility of cognition being scaffolded not only by physical objects in the environment but also social structures such as norms or conventions. Such science fiction staples as collective memory and group minds might be possible if vehicle externalism is true (see Wilson 2005), although neither are entailed by it. Vehicle externalism arguably does not entail content externalism (see Wilson 2004, p.179, for defence of this point), although the two are at least highly congruent.

Mental contents and vehicles, then, may plausibly be individuated according to factors which lie outside of our heads as well as those within. While contents are abstract entities which are semantically evaluable (they may be true or false, and attitudes towards them may be justified or unjustified), vehicles are widely assumed to be physical objects or processes. While propositional attitudes are abstract, like their content, mental representations are assumed to be physical vehicles whose interactions make possible the sequences of propositional attitudes which are held to constitute reasoning and rational thought. Contents and vehicles, likewise thoughts and mental representations, exist at two quite distinct levels of explanation.

1.1.3 Levels of Explanation

The concept of levels of psychological explanation stems primarily from David Marr's work on the psychology of vision, in particular his 1982 book *Vision*. There he distinguishes three distinct levels of explanation for neuroscience:

- *The computational level*: what function(s) the system fulfils, specifically its inputs and outputs, and what constraints operate upon it.
- *The algorithmic level*: what representations the system uses and what processes it employs to build and manipulate the representations.
- *The implementational level*: how the system is physically realised (e.g. what neural structures and processes implement the system).

Marr makes many detailed proposals about these levels, as part of a highly detailed and specific account of vision. The details need not concern us here; what has endured and permeated the study of cognition is the concept of engaging with the topic on multiple levels of explanation. While Marr sought to understand the brain using computational modelling, many contemporary philosophers of mind – and advocates of RTM in particular – seek to use comparable methods to understand the mind. Rather than Marr's three original levels, the literature on intentionality and mental representation typically makes use of two levels:

- *The personal level*: comparable to Marr's computational level, this is the level of conscious (and perhaps unconscious) thoughts. At least some of what occurs on this level is available to introspection.
- *The sub-personal level*: combining Marr's algorithmic and implementational levels, this level addresses the mechanisms and processes which make thought on the personal level possible.

Assuming either exist at all, then folk psychology occurs at the personal level and mental representations at the sub-personal. This can be clearly seen in Fodor's early definition of RTM:

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

Claim 1 effectively states that propositional attitudes require mental representations, and are produced by them. Claim 2 states that folk psychological reasoning is produced by sequences of mental representations, and the relation between the representations is causal. The use of 'tokenings' is significant; individual intentional thoughts exist at specific times and are linked to specific objects, yet comparable thoughts may be entertained at any time or place. Representational tokens are mental particulars which possess the features described above, i.e. are surrogates for external objects based upon their internal structure and act as vehicles for propositional attitudes. Representational types (often called concepts in the literature) are abstract entities with suitable structure to be instantiated as token representations. In this way the similarity between different thoughts can be rationalised with the existence of distinct mental representations for every instance of thinking.

1.2 Representational Theories of Mind

Let's summarise our discussion so far. RTMs posit the existence of a class of mental particulars called representations which explain the intentionality, or aboutness, of everyday thoughts. Individual representations act as surrogates for objects being thought about, in virtue of their internal structure, and hence make possible individual propositional attitudes which bear mental content. However it is individuated, content is abstract and semantically evaluable, and is what differentiates different examples of the same propositional attitude. Belief that *P* is distinct from belief that *Q* where *P* and *Q* are examples of different mental content. The sequences of propositional attitudes which, according to folk psychology, constitute our everyday thoughts are made possible by the existence of parallel sequences of mental representations which exist at a sub-personal level of explanation. While

propositional attitudes are the bearers of content, mental representations are the vehicles of content.

In the following section we will turn to how leading advocates of RTM develop this theoretical framework to deal with two principal problems it presents. Beforehand, a quick note on our characterisation of RTM so far. Linking mental representations to content in the way we have is quite uncontentious, similarly our use of levels of explanation. For example, Crane writes 'in current philosophy, the problem of mental representation is often expressed as 'what is it for a mental state to have content?'' (1995, p.25). Our presentation above is similar to, and informed by, Bermudez' *An Introduction to the Philosophy of Psychology* where he presents RTM as a functionalist theory of mind with the addition of internal structures for functional mental states. The main difference is that Bermudez restricts his characterisation to Fodor alone, while we have a broader remit. RTMs are enormously popular, and admit considerable variation. Our definition is designed to be as inclusive as possible, and as a result does omit some relevant distinctions. Some will be considered below, others will not. It is also possible that some who consider their theories to be RTMs will disagree with our definition. If that is the case, the arguments presented against RTMs will only apply to the degree that our definition applies to their theory. In the next section we will examine in detail leading examples of RTMs which fit our definition precisely.

1.2.1 Causal Semantics: Dretske, Millikan, Fodor

Earlier we described RTMs as being faced with two principal challenges to their goal of explaining intentionality: giving an account of thought about absent or non-existent targets, and giving an explanation in purely physical terms. The first problem arises because the structure of mental representations is supposed to be isomorphic to that of their targets (in veridical cases) or their objects (in non-veridical cases). How can a representation be isomorphic to something that does not exist, and how can we explain how and why mistakes

happen? The second problem arises from the explicitly physicalist character of RTMs.[5] Fodor gives a clear and fairly typical statement of the view:

> If the semantic and intentional are real properties of things, it must be in virtue of their identity with ... properties that are neither intentional nor semantic. If aboutness is real, it must be really something else.
>
> Fodor 1987, p. 97

Aside from being physicalists, the advocates of RTMs we will consider are also naturalists meaning that (to a first approximation) there is no discontinuity between the natural world investigated by physics, biology and chemistry and everything else. We will return to naturalism and its significance in chapter two.

Throughout this thesis we will be using Millikan and Fodor's RTMs as exemplars of a broader class of similar views. Restricting ourselves in this way is a pragmatic decision: the literature is vast, and to consider all significant variations would be utterly impractical in the space available. Millikan and Fodor are leaders in the field, and their theories stand out as the most detailed and sophisticated available, so it is entirely fitting for a critique of RTM to engage first and foremost with their views.[6] As mentioned above, I take it that the arguments against RTM developed later in this thesis will apply *mutatis mutandis* to other examples, but in the main they will be specifically applied to our two exemplars.

Millikan and Fodor are best illustrated by contrast with Dretske's theory, and so we will turn first to Dretske to act as a foil for the others. All three are causal theories, meaning that they aim to explain intentionality in terms of causal relations between physical objects. Since they are specifically concerned with explaining how mental states acquire their meaning, these theories are typically referred to as causal semantics.

[5] A dualistic RTM would, in my opinion, be ill-motivated but as far as I know there would be no contradiction or inconsistency in adopting such a theory, dependent of course on how it is cashed out. While many dualist philosophers make use of the term 'mental representation' I know of none who would commit themselves fully to the strong reading we are concerned with.

[6] For a somewhat broader overview of causal theories, see e.g. Adams & Aizawa.

A very crude version of the causal theory goes like this: thoughts are caused by objects in the environment, and for any thought we can identify a causal chain linking it back to its object. Mental representations acquire their content through just this causal connection. For visual perception, the causal chain would be light rays reflecting from an object's surface, which then stimulate receptors on the retina, which send signals along the optic nerve, which are then processed and so on.

Clearly this crude theory is inadequate for thought about non-existent objects since they cannot be participate in such causal connections. To borrow Harman's example (1990, p.34), Ponce de Lyon may have searched for the fountain of youth, but his search cannot have been *caused* by the fountain, as it does not exist. Unfortunately neither can the crude theory explain mistaken thoughts; as Fodor (1990) has argued, it renders all representational content disjunctive and hence indeterminate. The argument is this: if a thought about X can be caused not only by X itself in the veridical case, but also some other object Y in a non-veridical case, then there is no way of determining whether the content of that thought is X or $X \vee Y$. Clearly this generalises so that Y may be replaced with any disjunction of further terms – there is no limit to the ways in which we may be mistaken. We can, I think, agree with Fodor that addressing this disjunctive problem is a necessary requirement for any causal theory of mind, and clearly the crude theory fails to do so. Although it predates Fodor's disjunctive problem, Dretske's theory can be usefully seen as an attempted solution to it.

1.3 Dretske: Natural Information

Drawing inspiration from communication theory – the mathematical study of information – Dretske (1981) aims to provide an account of how our thoughts represent (and misrepresent) the things we think about by way of an account of the information they contain. He follows Grice's (1957) distinction between natural and non-natural meaning, taking the two to be closely analogous (though not identical). The key to understanding both

according to Dretske is information, and the way that information is transmitted in intentional thought can be understood by examining examples of natural meaning.

Information in the sense used by Dretske is ubiquitous – it is all around us – and is the raw material from which meaning is produced. 'In the beginning there was information. The word came later.' (Dretske 1981, vii). It is transmitted by means of signals which carry information about their sources; examples include the natural case of rings in a tree trunk conveying information about levels of rainfall across the seasons the tree survived, and the non-natural case of a car's fuel gauge conveying information about the amount of fuel in its tank. Reliable law-like variation between the size of the rings in a tree trunk and rainfall levels ensure that the rings carry information about rainfall levels. Similarly, law-like variation between the needle on a fuel gauge and the amount of fuel in the tank is necessary for the gauge to carries information about the amount of fuel present. If the fuel gauge is stuck the position of the needle no longer carries the same information, even if the amount of fuel in the tank corresponds precisely to the reading on the needle. Likewise a stopped clock never tells the right time, even on the two occasions a day when its hands are in the correct positions. There has to be a reliable *causal* connection (or 'channel', in Dretske's terms) for information to be conveyed.

Individual signals may, and in most cases do, carry multiple different pieces of information. At minimum, if the position of the needle carries the information that the fuel tank is half full, it also carries the information that there is a fuel tank, that either the fuel tank is half full or it isn't, and any other information logically entailed by the tank being half full. Dretske realises that there is no informational criterion for privileging one candidate over another: '[n]o single piece of information is entitled to the status of *the* informational content of the signal.' (1981, p.72). But if we are to ultimately explain the fixation of the content of propositional attitudes in terms of information – and Dretske is explicit that this is his aim – we must have some means of picking out *one* piece of information as providing meaningful content. Either that, or accept that intentional content is indeterminate, which would entail that there is no fact of the matter about what we are thinking. Although formulated slightly differently, we can clearly see Fodor's disjunctive problem at play here.

1.3.1 Determining Content

Dretske's solution is to impose a non-informational restriction upon what information gives a propositional attitude its content. In cases of conventional meaning this restriction is produced by stipulation. The position of the needle at 'E' means empty because that's how engineers design cars. Similarly, on a map the symbols (including lines) which are used are assigned their meaning by cartographers – they don't represent some terrain or location solely by merit of resembling it, there must also be a 'mechanism for embedding … [suitable] … information in the pattern of marks on paper' (1981, p.192). Strictly speaking, it is symbol types whose meanings are assigned in this way, with the meaning of token symbols on particular maps being derived from them. This reflects the fact that we don't need to have participated in the original baptism(s) of symbols to draw a map, or to read one accurately. There is a clear parallel here with causal theories of reference, where the referent of a term is fixed by an initial baptism, and is propagated by a causal chain of usage.

Essentially the same solution is applied to propositional attitudes (Dretske concentrates almost exclusively on beliefs, but we may generalise the point). There is no overt ceremony which baptises the meaning of thoughts, but Dretske hypothesises a learning period during which numerous causal signals are received. If during this period the cognitive system receiving all the signals develops a way to prune all the surplus information beyond, say, what tells us that something is *F* (in Dretske's terminology, to digitalise an analogue – that is perceptual – representation), then 'a certain type of internal state evolves which is selectively sensitive to the information that *s* is *F*.' (1981, p.193). By the end of the learning period these sorts of state types – concepts – have a fixed meaning by merit of developing an appropriate structure (*ibid.*, p.195). In the future, representations possess their content not in virtue of carrying a particular piece of information – after all, they carry many pieces – but in virtue of being a token of a particular type. A belief that *P* carries the content *P* in virtue of instantiating an appropriate physical structure (*ibid.*, p.213). This is analogous to a group of cartographers deciding to use blue lines to represent rivers: from that point onwards blue lines on maps *mean* rivers, even if they do not actually correspond to the

correct locations *of* rivers. Once a concept is acquired, it may be used to believe that something is the case regardless of whether there is a signal which carries to us the information that it is so. This explains both how informational content may be determinate, and how it may be mistaken.

In *Explaining Behaviour* Dretske develops this informational account of mental content in a number of ways, most significantly for our purposes by supplementing his account of content fixation with the teleological element of a given function. He introduces the term 'representational system', meaning 'any system whose function it is to indicate how things stand with respect to some other object, condition or magnitude' (1988, p.52), 'indication' being an alternative way of parsing the earlier claims about informational content. Representational systems then fall into three types. The first are purely conventional systems, ones which have no intrinsic powers of indication and which derive their functions entirely from ourselves, with some examples being maps, diagrams, musical notation, and prearranged signals.[7] The clearest cases involve ostensive definitions, such as using coins and popcorn to represent basketball players, where the intrinsic properties of the former have little if any connection to the latter (1988, pp.52-3). They represent the players purely in virtue of being given the function of doing so.

The second type of representational systems combines conventional and non-conventional, or natural, elements. Like the first type they are dependent upon us for the functions which determine what they represent, but unlike the first type they indicate how things stand independently of us and our uses of them. We give type two systems their role as representations, but they are roles to which they are already well suited – they already carry appropriate information for the job. Footprints in the snow are systems of this sort: they always indicate the movements of their maker, amongst other things, but only *represent* those movements when bequeathed the function of doing so.

[7] Some maps and diagrams do use conventionally defined symbols which also resemble the objects they represent – such as a square and cross for a church – so might perhaps be better treated as borderline cases with the second type. Which category individual non-intentional representational systems fall into needn't bother us greatly, however, as it is tangential to our concerns.

Type three systems, which include cognitive systems such as beliefs (1988, p.72), possess intrinsically both the ability to indicate how things stand (i.e. they carry information) and the functions which determine what they represent, which 'derive from the way the indicators are developed and used *by the system of which they are a part*.' (1988, p.62). That type three systems develop their own representational functions is the key move here, intended to make the leap from so-called 'derived' to 'original' intentionality whilst providing a substantial degree of continuity between natural and non-natural cases of meaning. As we shall soon see Dretske's teleological requirement that content is linked to function is applied somewhat differently by Millikan, and rejected by Fodor entirely.

1.3.2 Vertical and Horizontal Relations

While *Explaining Behaviour* develops Dretske's account of content fixation, both his later and earlier accounts incorporate essentially the same view of the 'vertical' relations between representations and propositional attitudes. The term vertical refers to relationships between members of the different levels of explanation discussed earlier, the personal level being 'above' the sub-personal. By contrast 'horizontal' relations are those which hold between members of the same level of explanation.

During a learning period of concept formation the thinker is exposed to a variety of signals indicating what does and does not bear some property, and develops a selective process to distinguish them. Although in later writings Dretske moves away from the concept of a learning period as such, he continues to emphasise the centrality of a historical basis for content fixation (see e.g. his 2006). At some point 'an internal structure is developed with the appropriate semantic content' (1981, p.195) and beliefs are identified with subsequent tokens of these semantic structures. So to believe that something is the case is to token a semantic structure in one's brain: 'beliefs … have been identified with physical states or structures. A belief is realised *by* or *in* a physical structure.' (1981, p.213) To this Dretske also adds the condition that the structure tokened should have a control function, which distinguishes beliefs from non-directive contentful states:

> A semantic structure qualifies as a cognitive structure (and therefore, we shall argue, as a belief) insofar as its semantic content is a causal determinant of output in the system in which it occurs.
>
> Dretske 1981, p.199

> What you believe is relevant to what you do because beliefs are precisely those internal structures that have acquired control over output … in virtue of what they … indicate about external conditions.
>
> Dretske 1988, p.84

Dretske has little to say directly about the horizontal relations between representations themselves, but we may indirectly derive some requirements from his discussion of simple and complex concepts (1981, pp.215-222). Simple concepts are used to express unitary ideas such as being red, whereas complex concepts are used for cases such as being a right-angled quadrilateral, which can be separated into the concepts of having sides, of having four of them, and each corner being a right-angle (*ibid*., pp.216-7). To this we might also add that further components such as each side being straight, meeting another at each of its ends, and none overlapping. The relationship between simple and complex concepts is, as this example suggests, compositional – complex concepts are *composed*, or made up of simple ones (*ibid*., pp.215-8). How this works – what process or processes combine concepts – is not explored; Dretske tentatively endorses Fodor's (1975) language of thought hypothesis according to which the syntactic operations between representations are linguistic in character, but says that the 'extremely sophisticated' mechanisms required are 'too technical' to discuss (*ibid*., p.230). We will return to compositionality and the language of thought shortly when we turn to Fodor.

So to believe that some thing *s* has the property *F* is, on Dretske's view, to have inside you (i) a physical structure, which is (ii) of a type whose earlier tokens indicated (carried the information) that *s* is *F*, and (iii) directs what we do. Their representational content is derived not from their own informational content (although true beliefs' informational contents will include ones which correspond to their representational contents), but from the content of the concepts they are tokens of. This is the essence of Dretske's account of misrepresentation – we may mistakenly believe that we see a cow when in fact it is a horse in front of us precisely because the representational content of beliefs is dependent upon

which concept they token, and not upon what actually causes that particular tokening. Put another way, misrepresentation occurs when natural meaning (via information) and non-natural meaning (via a system's function) are misaligned (Dretske 1986).

1.3.3 Indeterminacy of Content?

Dretske's theory of the semantics of intentional thought addresses the disjunctive problem and the issue of thought about non-existent targets in the same way: the content of individual mental representations is determined not by any direct relation with their targets or objects, but by their internal structure. The content of a representation is determined at a prior time when the meaning of all representations with that structure is established. Whilst information is never false, it is possible to be mistaken by tokening the wrong sort of structure. Likewise, we may think about something which does not exist since we have previously encountered other objects which bear the same properties, and formed appropriate concepts at that time. This applies not only to objects which do not currently exist, or are far away, such as tomorrow's dinner, but also those which never exist or even cannot do so. A representation of a unicorn may be composed from representations of horse and horn, suitably arranged; likewise a square circle.

However, as Dretske himself has observed, his account fails to explain how we can represent properties we have never encountered; his own example being 'dead' when applied to people. If there has never been an encounter with something similar, Dretske's account gives no way of explaining how representational content is fixed.[8] Further problems abound, many of which are discussed in McLaughlin (ed.) with replies from Dretske. As we are using Dretske as a foil to explain Millikan and Fodor we need not explore them all here. We will limit ourselves to one which is germane to our purposes.

[8] This is reminiscent of Descartes' painter analogy in his *First Meditation*, which questions the epistemic basis of our knowledge of composite objects.

There is a well-known criticism, pressed by Fodor (1990, p.70) amongst others, that teleological theories of content – that is, those which appeal to a concept of function to explain how representational content is fixed – fail to provide any determinate content. To take the classic example, when a fly passes across the visual field of a frog, the frog's tongue shoots out to catch the fly. Naturally, we might conclude that the function of this process is to catch flies; hence the state of the process at that time is about flies. But any small dark object passing across the frog's field of vision prompts the same response, and for many years researchers have been exploiting this by feeding frogs ball-bearings. In this case the frog's state seems to be about small dark objects, or perhaps about ball-bearings.

The argument is that in this case, and by extension in general, there is no fact of the matter which determines which interpretation is the correct one in general. Evolutionary theory tells us that since frogs usually live in environments where small dark objects are flies, they have no need to develop a function which distinguishes the two, and hence have not done so. An evolutionary account of function, or indeed any account which determines content based upon history as Dretske's does, cannot provide determinate content.

Dretske (1990) gives a stimulus-based response to this problem, according to which if an indicator *X* indicates two things, and it indicates the first in virtue of indicating the second, then *X* acquires the function of indicating the first. Let's use Dretske's example of a magnetosome possessed by some marine bacteria, which is sensitive to magnetic fields and steers the bacteria away from oxygen-rich water. Dretske's stimulus-based interpretation is that the magnetosome represents the direction of geomagnetic north, not the direction of oxygen-free water. This is despite the purpose of the magnetosome being to direct the bacteria away from oxygen-rich water which would be lethal. Similarly, the frog's visual process represents its stimulus of small black object, despite the purpose of doing so being to catch flies. Whatever its merits, this provides a determinate content for any process. Millikan, who also endorses a teleological theory of content, takes the opposite view to Dretske.

1.4 Millikan – Biosemantics

Like Dretske, Millikan's theory of content fixation is teleological, being based upon a concept of function, but she employs an explicitly biological concept of function to explain how this occurs. In contrast to Dretske, Millikan emphasises the role of the consumer rather than stimulus in determining content, and again in contrast her theory is disjunctive according to whether a thought is veridical or non-veridical.

For cases of veridical thought Millikan favours an informational account comparable to Dretske's above. While there are differences, such as Millikan's 'local' information including statistical facts and reference to individuals (see e.g. her 2004, ch.3), they need not concern us here as the essential framework is the same: information is ubiquitous in the environment, and veridical thought takes place through filtering and making use of it. The way in which thoughts represent objects is analogous to (and for Millikan, continuous with) natural signs such as the rings in a tree. The content of intentional representations is derived from naturally available information.

While Millikan's account of natural signs does vary from Dretske's in a number of details (in particular, her use of 'locally recurring' natural signs, see 2004 ch.4), with one exception the differences are not significant for our purposes. The significant difference is that Millikan takes representational content to be determined by the consumer rather than stimulus of an information signal, i.e. meaning is dependent upon use (see e.g. 1995, p.86). This will be central to her account of misrepresentation.

For cases of non-veridical thought, Millikan favours a very different type of explanation of content fixation to Dretske. Her strategy is to regard teleological theories as not being theories of content at all (2004, p.63), but explanations of misrepresentation alone. Thus her theory is disjunctive in character: veridical thought is explained informationally, and non-veridical thought teleologically in terms of normal or 'proper' biological functions.

1.4.1 Proper Functions

Consider the function of a coffee grinder (Millikan uses this analogy in her 2008). Coffee beans go in, the grinder grinds them, and ground coffee comes out. The proper function of a coffee grinder is to produce ground coffee, because this is what it normally does. When given the appropriate input of coffee beans, in the past it has reliably produced ground coffee, and we can reasonably expect it to do the same in the future. By comparison, the frog's strike mechanism has the function of catching flies because this is what it normally does. The purpose of the mechanism – and the evolutionary reason for its development – is to catch flies, not small black objects. The key here is the usefulness of the mechanism to the frog itself; catching flies aids survival, hence that is the function of the strike mechanism. In an environment where all small black objects are flies, on Millikan's view there is no need to invoke proper functions as the frog's strike is a natural sign of flies. In the same way, a magnetosome's direction is a natural sign of the direction of lesser oxygen – not geomagnetic north (2004, pp.44-5).[9] When something goes wrong, and a frog eats a ball-bearing or a magnetosome swims into oxygen-rich water, Millikan attributes the mistake not to a malfunctioning of the mechanism (although this, of course, can also happen), but to a change in features of the environment. In a laboratory setting the frog's strike mechanism functions normally but is presented with the wrong sort of stimuli to fulfil its purpose. When led astray by a magnet, the magnetosome functions normally but fails to fulfil its purpose of avoiding oxygen-rich water. If a coffee grinder is filled with cardamom pods, we can hardly expect it to produce coffee. The point is that in all of these cases, the failure of a mechanism to produce the best result is not attributable to any failure of that mechanism, but to changes in its environment. The function of a given mechanism is also determinate, being determined by its unique and specific proper function.

[9] To use Millikan's terminology fully, these are both examples of locally recurring signs, meaning to a first approximation that they make use of information with a reference domain extended across a portion of space and time, and they take the same meaning reliably.

1.4.2 Function and Intentionality

By analogy with natural examples such as those above, Millikan extends her biological teleological semantics – biosemantics for short – to intentional thought. Intentional signs, i.e. mental representations, are natural ones which have been selected-for, meaning that employing them *causes* enhanced survivability and/or reproduction.[10] Combined with Millikan's requirement of use determining content, this results in the view that intentional thought arose through the evolution of two symbiotic systems; one which produces signs and one which uses them (2004, p. 73). In cases of veridical thought the content of a mental representation is determined by its use by the representation consumer (which is either a person or some process in their mind), but will correspond to the available local information, i.e. to what is actually the case. Cases of non-veridical thought which are not caused by cognitive impairment such as through brain damage arise due to unusual factors in the environment.

Let's return to our example of seeing John in the distance, and mistakenly thinking of him as Tom. Here, on Millikan's view, our sign consuming system is functioning normally and this fixes the representation's content as being about Tom. The mistake lies in the sign's production, as somehow the presence of John has produced a sign which is normally caused by Tom; perhaps it is dark, or John is wearing Tom's clothes etc. What matters is that our *use* of the intentional representation is entirely normal, it is the context in which it is produced which has generated the mistake.

1.4.3 Horizontal and Vertical Relations

Given her pragmatic approach to content fixation, according to which the use a system or organism makes of a sign or representation determines its content, Millikan's construal of personal-level cognition is relatively broad. She draws numerous examples from animals and infants as well as adult humans, and the evolutionary foundation of her theory places a

[10] I take this definition of selection-for from Sober which Millikan explicitly references in her use of the phrase.

firm emphasis on actions and their results, in terms of enhancing survivability and reproduction. This might seem a far cry from the highly conceptual sphere of commonsense psychology and propositional attitudes, but Millikan sees all cases of cognition as being in a continuous spectrum. Whether intentional or otherwise, representations admit of three main types: descriptive, directive, and both. Descriptive representations are those which are satisfied by corresponding to features in the world, while directive representations are satisfied by actions which cause the world to match their content. Representations which combine both functions are called pushmi-pullyus after the two-headed llama in Hugh Lofting's Doctor Doolittle stories. According to Millikan, pushmi-pullyus are the most primitive type of representation, and account for the majority of animal cognition (2004, p.80; see also her 1996). Descriptive and directive representations, such as beliefs and desires respectively, are more refined (and, for Millikan, less interesting) and are supposedly the preserve of the cognition of humans over the age of around 3 or 4.[11] It is these more refined representations which are intentional, both in the sense of being selected-for and in the sense of being *about* things, and which are linked to propositional attitudes:

> [I]ntentional representations always come with propositional attitudes attached ...
> There are not and could not be intentional representations that lacked attitude.
>
> Millikan 2004, p.81

> The content of the representation turns out to be an abstraction from a fuller affair intrinsically involving [a] ... propositional attitude.
>
> Millikan 2005, p.5 of online version.

In this way intentional representations are clearly distinct from non-intentional ones, while a close analogy is maintained with more primitive natural meaning through the dual nature of pushmi-pullyu representations. Personal-level cognition amongst adult humans is clearly framed in terms of propositional attitudes. Whether Millikan directly endorses commonsense psychology, the view that cognition primarily involves *explanation and prediction* in terms of propositional attitudes, is unclear – I have been unable to find any

[11] Millikan endorses the widely-held view, based upon the psychological experiment known as the False Belief task, that infants under the age of 3 or 4 are literally unable to believe anything, on the grounds that they fail to ascribe correct beliefs to others under some circumstances (see her 2004, ch.1). We will examine – and reject – the empirical basis for this claim in chapter five.

mention of it in her work at all, but her commitment to propositional attitudes is beyond doubt.

Millikan's view of horizontal relations between representations, which occur at the sub-personal level, concerns their status as structured entities. Representations have a 'compositional semantics' (1995, p.90) meaning that their meaning is reflected in their parts plus the arrangement of those parts. Although the content of a representation is determined by the use to which it is put, 'mental representations are systems of brain happenings or brain states that map onto represented world affairs' (2004, p.84), and they do this in virtue of their constituent structure. *It is this structure which makes representations useful* and which makes possible the isomorphism, a close structural correspondence, between representations and their targets which is a central tenet of all varieties of RTM. In Millikan's terms, the relation is an 'isomorphism, in the abstract mathematical sense, between the domain of the signs and the domain of the signifieds' (*ibid*.)

The composite structure of representations is central to their horizontal relations. Millikan is deliberately vague about the precise nature of this compositionality – in contrast to Fodor who, as we shall see, is very explicit on this point – sometimes making use of analogies such as maps and pictures, whilst failing to endorse these as literal claims. In one such example from her 1995, Millikan describes the process of orienteering by combining two representations which, like incomplete maps, each contain some local geographical features and their relative locations. It is the partial overlapping of the symbols in these map-like representations which allows them to be synthesised into a representation which includes all the information from both, and it is this representation which is useful to the organism in question. Whatever the nature of the way in which representations are codified – and they always are (2004, p.84) – their internal structure is central to the ways in which they can interact.

It might seem odd that Millikan is so vague about the way or ways in which representations are codified, and hence their internal structure, but once again her pragmatic approach to content fixation underlies her view. Just so long as the consumer of a representation is able

to make use of it effectively, it really doesn't matter how it is encoded. It is true that simplicity is generally preferable, but not at the expense of effectiveness (*ibid*.). Also, given that Millikan takes representations to be brain states or events, from her biosemantics it follows that determining their structure is likely to be a largely empirical affair.

1.4.4 Biosemantics and Misrepresentation

To summarise, Millikan adopts an informational theory similar to Dretske's to explain how content is fixed in cases of veridical thought. Where the target of thought is absent, Millikan claims the mental process involved still functions as it would to produce a veridical thought if it were in normal circumstances. It is the abnormal circumstances which produce the mistake. This applies equally to thoughts about non-existent or impossible targets, as 'circumstances' is not limited to the physical environment alone. Representations arise from interplay between the systems which produce them and which consume them, with their content being determined by their use but reflected in their constituent structure. It is the isomorphism between mental representations and the features of the world they stand in for which make them useful to us.

Millikan's theory does seem to provide a solution to the disjunctive problem, in that it identifies a specific determinate content for any representation, according to the way that representation is used. On the assumption that it is possible to give a unique specification for every use of a representation, then every item of content will also be unique.

1.4.5 Problems for Millikan

There are a number of objections to Millikan's biosemantics, although we need not canvass them in detail here. Firstly, there are a number of good discussions already present (see e.g. Adams and Aizawa section 3.2.1; also Cummins 1996 for some sustained criticism of her earlier formulations); secondly our purpose here is not to criticise Millikan's theory on

semantic grounds. Recall that after exploring the nature of RTMs, our project will be to engage with them as ontological theories, and as such our lines of criticism will be quite distinct from those which object to their semantics. One objection which is worth mentioning is due to Fodor 1990b, who questions whether natural selection acts at the level of individual representations. The concept seems admittedly odd, especially given the degree of variation between people's brains (which representations are supposed to be coded within), but if entertaining particular thoughts can causally increase one's survivability or reproductive chances, and it seems plausible that this is the case, then individual representations would satisfy the definition of selection-for, and hence be suitable for natural selection.

1.5 Fodor – Asymmetric Dependency and the Language of Thought

Like Dretske and Millikan, Fodor aims to give a naturalistic explanation of intentionality, one which makes no reference to intentional concepts (see e.g. Fodor 1987, p.97), likewise an explanation which is explicitly causal. Unlike those other theories, Fodor's is not teleological: he makes no use of functions or historical facts. To solve the disjunction problem, and explain misrepresentation in general, Fodor uses the concept of asymmetric dependency. The basic idea is simple: the existence of non-veridical thoughts (those with incorrect content) is causally dependent upon the existence of veridical thoughts (those with correct content), and not vice-versa. While details have varied somewhat across different versions of the theory (Fodor 1987, 1990a, 1994), the core of the theory is that a content 'X' means X when:

1. "Xs cause 'X's" is a law.
2. For all Ys that are not Xs, if Ys qua Ys actually cause "X"s, then the Y's causing "X"s is asymmetrically dependent on the Xs causing "X"s.[12]
3. The dependence in (2) is synchronic.

[12] If we are to be pedantic, the dependency relation here is antisymmetric since it is symmetrical in the reflexive case of Xs causing Xs. For clarity's sake we will use Fodor's 'asymmetry'.

To use one of Fodor's examples, imagine mistaking a cow for a horse. There is a law to the effect that horses cause the content 'horse', meaning that this is what usually happens (for present purposes 'law' can equally be interpreted as 'law-like correlation'). If we see a cow, and *as the object of thought* that cow causes us to entertain a thought about it with the content 'horse', then the cow causing 'horse' is dependent upon horses causing 'horse'.[13] Horses causing 'horse' is not dependent upon cows – or anything else which isn't a horse – causing 'horse'. Lastly, the asymmetric dependency is synchronic, meaning it occurs at that time rather than across times – this third principle serves to rule out some pathological cases, and also emphasises the contrast between Fodor's theory and Dretske and Millikan's which take content fixation to be necessarily diachronic.

Asymmetric dependency seems to capture an important truth about non-veridical thoughts, namely that they are effectively parasitic upon veridical ones. Mistakes are only possible against a background of largely correct thoughts, in much the same way that lying is only possible in a context of telling the truth. As Kant observed, universal lying would be self-destructive because it would undermine the very concept of truth which lying is dependent upon; likewise, all thoughts could not be mistaken as this would sever the relationship between thoughts and their targets which is necessary for the concept of veridicality in the first place. Asymmetric dependency also seems to deliver the correct result in examples of the type we have used: 'horse' thoughts caused by cows are possible precisely because 'horse' thoughts are usually caused by horses. In this case, for whatever reason the cow was misidentified. Similarly our previous example with John and Tom can be explained: contents produced when John is the object of thought and Tom the target are dependent upon the veridical case of Tom being both target and object. Thus asymmetric dependency can account for misrepresentation occurring where the target of thought is absent.

The disjunctive problem is similarly disposed of: while 'horse' can be caused by horses or by cows, and hence by horses or cows, Fodor's theory provides a unique specific content for every thought, which is the content produced in the law-like case stipulated in (1) above. If there is a law to the effect that Xs cause 'X', then that law determines the content of a representation, and the *actual* cause of the representation is irrelevant, provided that the

[13] This qualification is important, in order to rule out odd scenarios where an object causes a thought through some non-standard means such as a blow to the head.

cause of the content is asymmetrically dependent upon the law in question. To put it another way, we can mistake anything for anything else; the content of our thoughts is based upon what reliably and routinely causes that content, not whatever happens to cause it in any one situation.

Thought about non-existent objects is dealt with separately. There can be no law that says Xs cause 'X' where Xs do not or cannot exist. Fodor is a semantic referentialist, meaning that he rejects the existence of Fregean senses, which are widely considered to exist based upon beliefs producing so-called opaque contexts. The issue is this: 'X believes that Y is Z' may be true while 'X believes that W is Z' may be false, even though Y and W are identical. So although the morning star actually is the evening star, it does not follow that what is true of one is true of the other in such cases. The Fregean explanation is to distinguish between the reference of these coextensive terms – which is the same in all cases – and the senses of the terms which may be different. Fodor rejects senses, differentiating coextensive terms by the way they are structured. We will return to Fodor's argument for this view in chapter six, but in brief his solution depends upon conceptual atomism, according to which the fundamental building blocks of reference are singular terms and predicates, and compositionality, which claims that the meaning of complex concepts is determined by the meaning of their parts plus how they are arranged. We will return to compositionality in discussing horizontal relations between representations.

For now, we can see how these three principles explain thought about non-existent objects. Similarly to Dretske, by compositionality Fodor regards complex concepts such as unicorn to be made up of simple ones such as horse and horn plus their manner of arrangement. By atomism it is these simple concepts which bear meanings directly, and by referentialism they derive their meaning directly from the objects they refer to. Since horns and horses *do* exist, there can be laws about the contents they cause, and hence thought about things which do not exist is grounded in thought about things which do. The same story applies to things which cannot exist, and since Fodor is committed to some concepts being innate (i.e. not acquired through experience, see e.g. his 1975, 1983, 1990a), it can also apply to simple

non-existent objects and complexes composed in part or full of simples which are non-existent, provided concepts of the non-existent simples are innate.[14]

1.5.1 Problems for Asymmetric Dependency

There are naturally a range of well-known objections to Fodor's theory; once more, we need not canvass them here, but one less well-discussed is worth mentioning to help clarify how the theory works. Asymmetric dependency can be regarded not as a theory *per se*, but as a theory schema. In much the same way that RTM admits many variations which preserve certain key features, there could be a wide variety of different relations which satisfy Fodor's three requirements above. There are many ways in which one thing can be dependent upon another, and many ways in which one thing can be *causally* dependent upon another as well. Asymmetry is a property of many binary relations, and in no way specifies any unique one. Fodor's schema requires a dependency relation which satisfies the requirements of being causal and asymmetric, but there are plausibly multiple candidates which could fulfil this role.

This in itself is not necessarily an objection to Fodor, though it does raise two concerns. The first is that Fodor's theory is not on an equal footing with Dretske and Millikan's, which both specify a single relation which determines the content of representations. For Millikan this is the use of a representation based upon its being selected-for, for Dretske the relation is less clearly specified but is based upon a 'certain type of learning' (Dretske 2006, p.72) having occurred in the past. Thus it might be said that those theories are more precise than Fodor's schema, and a direct comparison would only be possible if more were said about the nature of the asymmetric dependency Fodor makes use of. On this line of thought, there is currently a lacuna in Fodor's theory which may or may not be easy to fill. The second concern is closely related, in that if satisfying Fodor's schema requires use of intentional concepts then it will have failed to provide the analysis of non-natural meaning in natural

[14] In his 1975 Fodor in fact argues that *all* concepts are innate; this is not a requirement of his theory, and seems to have been dropped in more recent writings.

terms which he explicitly aims for (essentially this point is made by Adams & Aizawa, section 3.4.1). Whether this is the case remains to be seen.

1.5.2 Horizontal Relations: Propositional Attitudes and the Language of Thought

Fodor's account of horizontal relations at the personal level is admirably clear and consistent. A self-confessed 'realist' about propositional attitudes, from his 1975 onwards, Fodor takes it that 'explaining actions by attributing beliefs and desires to their agent is the very paradigm of how a mentalistic psychology does its thing' (2008, p.5). Furthermore, beliefs and desires are discrete propositional attitudes which satisfy three conditions:

> (i) They are semantically evaluable.
> (ii) They have causal powers.
> (iii) The implicit generalizations of commonsense belief/desire psychology are largely true of them.
>
> In effect, I'm assuming that (i)-(iii) are the essential properties of the attitudes. This seems to me intuitively plausible.
>
> Fodor 1990, p.10

In effect, Fodor fully endorses the characterisation of commonsense psychology above. Patterns of thought such as reasoning and inference are explained in terms of sequences of propositional attitudes, with their principal functions being the prediction and explanation of behaviour. To understand why someone takes an umbrella outside with him, we simply need to understand it is because he believes it is raining and desires to stay dry.

There are some limitations to this claim, since in recent years Fodor has come to accept the existence of some mental content which is non-conceptual, and hence cannot be the content of a propositional attitude. His view, however remains that propositional attitudes are paradigmatic in the functioning of the mind, while non-conceptual content plays a relatively minor role (for details see e.g. his 2008, ch.6). Like Millikan, he characterises

personal-level thought in terms of propositional attitudes, though unlike her he admits some few exceptions and endorses a specific view of how propositional attitudes interact.

Fodor's account of horizontal relations at the sub-personal level is equally explicit. His language of thought hypothesis (LOT) states that thinking takes place in virtue of a symbolic system realised in the brain. This symbolic system has a language-like syntax, and is literally a language of thought (often called 'mentalese'). Precise formulations of this hypothesis have varied over the years, at times leading to complicated definitions (for a case in point see Aydede section 1); to avoid unnecessary complications we will limit ourselves to Fodor's most recent formulation. As the title suggests his 2008 book *LOT2: The Language of Thought Revisited* is devoted primarily to this theory, making a number of corrections and changes to earlier versions. One such is a massively increased emphasis on the compositionality of thought, which Fodor now views as absolutely central to cognition in general and LOT in particular.

Fodor's main argument for LOT is as follows: the single most important feature of thought for cognitive psychology to explain is its productivity, how it is that we can combine and recombine a finite number of concepts, connectives and so on to produce an effectively infinite range of thoughts. The answer, he claims, lies in the compositionality of thought, and compositionality entails LOT, amongst other things.[15] Compositionality is 'at the heart of the productivity and systematicity of thought' (ibid. p.20), systematicity meaning that the ability to think certain thoughts is linked to the ability to think others.

As mentioned earlier, compositionality is the view that representations are structured entities, with complex concepts being formed from simple ones plus their manner of arrangement. From this it follows that 'the content of a thought is entirely determined by its structure together with the content of its constituent concepts' (Fodor 2008, p.17). The complex concept brown cow derives its meaning from the simple concepts brown and cow plus their arrangement; rules of syntax determine which arrangements are well-formed, ruling out cow brown in this case. Thus representations have both a syntax and semantics

[15] Fodor also takes compositionality to entail semantic referentialism, since reference generates transparent contexts while senses do not (2008, p.20).

which is 'intimately dependent on their constituency' (2008, p.106). As language exhibits productivity, systematicity and compositionality, and has a definite syntax, it is the obvious model upon which to base thought.[16]

1.5.3 Computation versus Association

To summarise so far, at the personal level Fodor takes thought to consist in propositional attitudes which are themselves relations to token mental representations. Simple mental representations derive their meaning directly from the objects they refer to, or else are innate, and complex mental representations are structured entities whose meaning is derived from the meaning of their parts plus their structure. The manner in which mental representations are structured has a language-like syntax, and hence we literally think in a language of thought.

In parallel to Marr's third level of explanation – that of physical implementation – Fodor also endorses a view of how representations are implemented, which he calls the computational theory of mind, or CTM. This is the view that the reasoning processes which combine series of mental representations are sensitive only to their syntactic properties, and these processes can be modelled computationally. On this view the mind is analogous to a digital computer (strictly speaking, a Turing machine), with mental representations being symbols akin to those processed by a computer operating on so-called classical architecture. The finer details of CTM need not concern us here, as it is strictly speaking a separate theory from RTM (Fodor himself is quite clear on this point, in numerous places), but as part of Fodor's overall view of the mind it is helpful in illustrating his RTM, which it complements.

Fodor presents an idiosyncratic history of RTM according to which it was originally developed in broad terms by Aristotle (see Fodor's 2008, p.5), and endorsed by Descartes,

[16] In fact, my opinion is that despite being perhaps the most well-known of Fodor's views, his language of thought hypothesis is peripheral to his theory of mind. Fodor is correct in identifying compositionality as the key feature RTMs need to explain, but there is simply no need to assume linguistic composition. It is telling that while Fodor's 2008 book is titled *LOT2* in homage to his earlier work, the language of thought hypothesis barely gets a direct mention, and is not even referenced in the index.

Locke and Hume in varying forms, before being updated in the Twentieth Century by himself in particular. The central thesis of his book *Hume Variations* is (rightly or wrongly) that his own RTM is 'more or less interchangeable' (Fodor 2003, p.8 fn.2) with Hume's associationist theory of Ideas, only with the addition of internal structure for concepts. Fodor's main argument is that association is semantically transparent: the content of A-associated-with-B is just the content of A associated with the content of B. So associationism cannot distinguish between the single complex concept brown cow and the sequence of two concepts brown and cow. In contrast, a computational mechanism can distinguish between the two cases as it can model truth-functional operators, as used in electronic logic gates and in propositional logic. Thus, the conjunction of brown and cow, i.e. brown cow, can be distinguished computationally from its two elements.[17] Such a mechanism will have to be compositional, because it tracks the relationship between complex concepts and their constituent parts.

Fodor's interpretation of Hume is certainly questionable (likewise Locke, Descartes, Aristotle and so on), but his argument effectively illustrates the importance of compositionality in understanding the horizontal relationships between representations as both simple and complex structured entities. Whether they are implemented in a manner best modelled by computational processes is strictly a separate issue from RTM itself, so one on which we will take a neutral stance. Since our ultimate aim is to demonstrate mental representations do not exist, the manner in which they are implemented will be moot.

1.5.4 Vertical Relations

The vertical relations between mental representations and propositional attitudes vary little across different varieties of RTM. All varieties take sequences of thoughts at the personal level to be dependent upon sequences of symbolic mental representations at the sub-personal level, with individual propositional attitudes corresponding to token mental representations. The internal structure of each token mental representation is crucial for

[17] Given the associativity of conjunction, however, brown cow cannot be so readily distinguished from cow brown. Here the syntax of mentalese would be required (i.e. a return to RTM).

determining its content and *ipso facto* which attitude it is related to. Belief that *P* is only different to belief that *Q* where $P \neq Q$.

Differences do exist: Millikan takes a representation's use to be the principal factor determining its content (though recall it has an internal structure suitable for that use), while for Fodor content is determined by a combination of syntax and semantics. Millikan takes every representation to have a propositional attitude 'attached'; Fodor admits some exceptions (so-called preconceptual representation) whilst agreeing for everyday thoughts. Millikan is neutral regarding the way in which representations compose and are implemented in the brain, while Fodor takes their composition to be linguistic and endorses the separate view they are implemented in a computational system. Lastly, Millikan does not clearly define what propositional attitudes are – to be fair, their use is so widespread in philosophy of mind she probably feels no need to do so. Fodor defines propositional attitudes as 'relations between minds and token mental representations' (2008, p.7). Thus he writes 'to explain what it is for a mental representation to mean what it does *is* to explain what it is for a propositional attitude to have the content that it does.' (2003, pp.8-9).

1.6 Summary of Causal Theories of Mind

	Dretske	*Millikan*	*Fodor*
Personal level (for human adults and older children):	Propositional attitudes.	Propositional attitudes.	Propositional attitudes.
Sub-personal level:	Sequences of mental representations.	Sequences of mental representations.	Sequences of mental representations.
Representations are:	Mental particulars. Symbolic. Compositional. Encoded in a language of thought.	Mental particulars. Symbolic. Compositional. Encoded in some way, which is analogous to maps in at least some cases.	Mental particulars. Symbolic. Compositional. Encoded in a language of thought.
Implementation in the brain:	-	-	Computational Theory of Mind
Veridical content is fixed by:	Prior formation of a representational type, possibly during a learning period.	Use of the representation by a consumer.	Causal laws.
Non-veridical content is fixed by:	Mistakenly tokening a representation type which had been previously learned.	A mechanism performing its proper function in a non-standard environment.	Asymmetric dependence upon veridical content fixation.

Here we will leave Dretske, and concentrate solely on Millkan and Fodor, whose theories exemplify the most developed and sophisticated varieties of RTM at present. So far our presentation of RTM has been largely exegesis. In the following section we will justify the characteristically metaphysical approach taken in the remainder of this thesis.

2. Metaphysics and Mental Representation

As we have seen in our examination of RTM above, and as is borne out in the vast literature on RTM *tout court*, interest in the field has been almost exclusively semantic in character. By semantic we mean concerned with what it is for thoughts to have meaning (and in particular truth values), with RTMs exploring this issue through what it is for a thought to have content. Given the more-or-less universal views that meaning is explained by content, and that the problem of intentionality is that of explaining how thoughts are about objects (neither of which are limited to RTM), this emphasis on semantics – although it is almost to the exclusion of all else – is understandable, if not necessarily benign.

As mentioned above, the issue of how to explain the meaning of veridical and non-veridical thoughts is legitimate, and so the semantic theories we have considered are well-motivated since they address that problem. However, RTM is more than a purely semantic theory. It is also a metaphysical theory, more specifically an ontological theory as it makes a very definite claim about what sorts of things exist: RTM requires that there exists a class of entities which are internally structured mental particulars, and which satisfy the requirements made by Millikan, Fodor and others, i.e. which are mental representations. If it should turn out that mental representations do not exist, then *ipso facto* all varieties of RTM are false.

In a sense this point seems trivial, but it is worth making clear. RTMs do not employ mental representations in anything like the weak reading we mentioned at the start of chapter one, nor as a catch-all term or convenient fiction. RTMs explain the intentionality of thought as occurring in virtue of there existing token structures in the thinker's brain which have suitable relations between each other and the propositional attitudes which one is thinking. No mental representation means no intentional thought, and hence RTM is fully committed to the existence of mental representations.

2.1 Ontological Commitment

Unfortunately, what counts as a genuine ontological commitment is difficult to establish. It is unlikely that simply saying 'there is *X*' commits one to the existence of *X* as counterexamples abound. 'There is a lovely smile on her face' does not entail that there is an entity – the smile – over and above the face (see e.g. Varzi 2002 for many more examples). On the other hand, even if 'there is' statements are ontologically neutral, there are unquestionably times when they do express a commitment (see Azzouni 2007 pp.204-5 for a more detailed explanation of this point). Thus the commitment lies not in the language used, but elsewhere. A useful means of differentiating sentences which express ontological commitment from those which do not is paraphrase: if a different sentence without 'there is' (or an equivalent) serves equally well, then the original statement was not being used to express a commitment in the first place. 'There is a lovely smile on her face' easily translates into 'she is smiling in a lovely manner' which has no apparent commitment to smiles as distinct objects. Clearly the claims made about mental representations above cannot be readily paraphrased in this way. An adverbial attempt could be made using 'thinks representationally', but the specific details of Millikan and Fodor's theories demonstrate that representations must be construed as discrete entities to fulfil their horizontal and vertical relations. On this linguistically neutral view of ontological commitment, RTM is committed to the existence of mental representations.

Of course, there are other accounts of ontological commitment. It may be thought that literally every 'there is' statement expresses commitment. If that is the case, the commitment of RTM to mental representations would be indisputable. Alternatively Quine's (1951) criterion tells us to regiment an area of discourse – such as by rendering it in predicate logic – and then reading off its commitments according to each existential quantifier. This method has received significant criticism recently, with Azzouni (1998) observing that the regimentation process is unnecessary (see e.g. Raley pp.291-2 for further details), and it is difficult to establish how readily RTM could be regimented into a logical form. Certainly it would be a lengthy and complex task, and – when the need to do so is questionable – one which seems not to be worthwhile. Furthermore, RTM is in a certain

sense already highly regimented. Analysing all the richness and diversity of thought into a framework of tokening discrete mental particulars is very much a formal process – using formal in a rather specialised sense we will return to shortly. On either the Quinean view or Azzouni's modification of it, this plausibly makes RTM suitable to read ontological commitment from.

2.1.1 Ontological Naturalism

There is a possible means of avoiding the conclusion that RTMs are committed to the existence of mental representations. As mentioned above, the varieties of RTM we are considering (as well as many not directly addressed) are naturalistic. Naturalism is an imprecise term which admits a wide range of interpretations. Broadly speaking, however, we can identify three types:

> *Minimal naturalism*: there is no radical discontinuity between what is natural and anything else.
>
> *Methodological Naturalism*: philosophy and science should employ closely similar methods, with any differences being superficial.
>
> *Ontological Naturalism*: the only things which exist are those required by natural science.

Minimal naturalism is the standard view in analytic philosophy of mind and beyond, and any RTM which seeks to exploit an analogy between natural and non-natural meaning – as do all causal varieties including Dretske, Millikan, and Fodor – will necessarily be at least minimally naturalistic. This is the weakest position. Methodological naturalism is a stronger view, and ontological naturalism the strongest. Each position arguably entails the weaker version(s), though in practice there are nearly as many types of naturalism as there are naturalists; we have only picked out the three most popular. For an overview of the many different views see e.g. Papineau.

If advocates of RTM are also ontological naturalists, they may claim no commitment to the existence of mental representations on the ground that their ontology is dictated by natural science alone, and not by any purely philosophical or conceptual concerns. But is this claim actually made? As mentioned above, all causal varieties of RTM are clearly committed to at least minimal naturalism. Fodor directly rejects methodological naturalism: 'I'm after a theory of the cognitive mind. I don't care whether it's a philosophical theory or a psychological theory or both or neither, so long as it's true' (2008, p.22), and his methodological openness is reflected in recent discussions on topics such as varieties of composition (see e.g. his 2003 ch.2) and the metaphysics of reference (2008, ch.7), which employ decidedly non-empirical methods and argument. In rejecting methodological naturalism, Fodor effectively rejects ontological naturalism as well.

While Millikan is explicit in her endorsement of naturalism (see e.g. her 2000c p.1), I know of only one place where she discusses its meaning or implications. In *On Clear and Confused Ideas* she states her evolutionary approach to cognition has a

> methodological implication that should be kept constantly in mind. If we are dealing with biological phenomena, then we are working in an area where the natural divisions are divisions only de facto and are often irremediably vague. These divisions do not apply across possible worlds; they are not determined by necessary and/or sufficient conditions.
>
> Millikan 2000a, quote taken from online version

This is certainly compatible with methodological naturalism, though insufficient to conclude that Millikan definitely endorses that position. There is no evidence to suggest that Millikan endorses ontological naturalism. So neither Fodor nor Millikan can use ontological naturalism as a way to argue against their ontological commitment to mental representations.

But what of the many other RTMs which we take Fodor and Millikan to be exemplars of? If their advocates endorse ontological naturalism, then our claim that RTM is in general committed to the existence of mental representations is severely limited. In response to this we will make two observations. First, after a number of years reading the literature on mental representations, I have not encountered a single mention – much less endorsement

– of ontological naturalism in connection with RTM. This does not mean no such endorsement exists, but it is certainly not commonplace. Nor does it mean that advocates of RTM are not free to endorse ontological naturalism should they wish. However, and this is our second observation, ontological naturalism is a controversial philosophical position, and not one which can be assumed by default to be true. There are strong arguments to suggest that ontological naturalism is mistaken. For example Raley argues that we cannot straightforwardly read off ontology from science: 'we must decide which parts of scientific discourse are true, and (or) which criterion for ontological commitment is correct. Such philosophical manoeuvring undercuts the ontological naturalist's official strategy' (2005, pp.292-3). Ontological naturalism is self-defeating on this view, guilty of a kind of pragmatic mistake.

Regardless of whether Raley is correct in this analysis, ontological naturalism cannot be assumed to be acceptable or true without explicit defence. As this defence is entirely absent from the literature on RTM, it cannot be used to block that theory's commitment to the existence of mental representations.

2.2 Concerns and Methods in Ontology

In exploring RTM we have seen that the semantic approach which characterises most of the literature has its own particular concerns. These issues about meaning, specifically over how content is fixed in general and how mistakes can occur in particular, have shaped the debates over which variety of RTM is to be preferred, and have also influenced the methods used by philosophers to explore them. In engaging with RTM as an ontological thesis, one also concerned with what exists, we will discover a distinct set of concerns and methods. The aims of this chapter are to identify these concerns and methods, and to argue that they are not only available for the study of RTM but are absolutely necessary for a full understanding of the theory. In chapters three and four we will illustrate how the concerns and methods of ontology play out using the example of the part-whole relation, before applying the morals from that case to mental representation in chapter five.

There are three reasons for choosing the part-whole relation as a means of illustrating ontological concerns and method. The first is that in recent ontology, parthood is universally regarded as one of – if not the – fundamental concepts available, and is routinely used in the analysis of others. As Smith & Grenon observe, it is 'a generally accepted feature of recent work in ontology that the part relation is used as a crucial organising tool' (p.280). Secondly, the nature of the part-whole relation has been thoroughly explored across the Twentieth-Century, primarily through a theory known as classical mereology, and so provides plenty of valuable material to draw upon. Thirdly parthood is an especially relevant concept for the study of mental representation since as we have seen these representations are characteristically internally structured. This structure is a relationship which necessarily holds between parts and whole, so a full account of the ontology of mental representation will incorporate an account of parthood.[18] This point has been sadly overlooked in the literature on RTMs, with the exception of a passing reference by Fodor who observes that the compositionality central to RTM is dependent upon 'constituency, which is a mereological relation that holds between a complex representation and the constituents out of which it is constructed.' (2008, pp. 105-6, italics removed).[19]

RTMs should be considered from an ontological point of view since they are committed to the view that mental representations exist, and the concerns and methods of that point of view are best illustrated through analysis of the part-whole relation. There can be no reasonable doubt over whether this is a legitimate means of analysing RTM. We will shortly argue that it is also a necessary requirement of a proper analysis. To do so we need to understand the concept of so-called 'formal' ontology.

[18] In case we suspect it is only *mental* representations which are structured, other examples also show an eliminable reference to parthood. All representation is partial in the sense that the representation does not reflect every aspect of what is represented. A legal counsel represents your legal interests in court, not the colour of your underwear; a good legal counsel, anyway.

[19] Composition alone is plausibly insufficient, although necessary, to constitute a full ontology of mental representation. The manner of composition is certainly relevant, though as we have seen in chapter one only Fodor has anything significant to say on this score, by way of his language of thought hypothesis. Simons (in conversation) also suggests that Fregean saturation is a relevant concern. I am inclined to agree, though as this thesis is concerned with taking a first step in exploring the ontology of mental representation, rather than having the last word, I feel that addressing all possible relevant factors is outside its scope.

2.3 Formal Ontology

There is no single universally accepted definition of formal ontology, though it is by no means as broad a term as naturalism, for example. We will start with Lowe's definition, which can reasonably be considered a standard view:

> Formal ontology is that branch of analytical metaphysics whose business it is to identify ontological categories and the formal ontological relations that characteristically obtain between the members of different categories.
>
> 2006 p.69

Formal ontology then, to a first approximation at least, is concerned with categories and cross-categorial relations.[20] In less loaded terms, we might say that formal relations are perfectly general, or that they are topic-neutral. It should be clear enough that parthood fulfils this requirement; talk of parts and wholes is ubiquitous in literally every domain of discourse. Admittedly, there may be some categories or domains whose members do not actually enter into part-whole relationships either contingently or necessarily (perhaps either where every member has been destroyed, or else are some suitable abstract objects, respectively), but the point is not that formal relations need to be exemplified across every category. Rather, it is always a pertinent question whether or not they are. To ask where propositions are located, for example, may be to commit what Ryle called a 'category error', whereas to ask if they have or can be parts certainly is not – even if it (improbably in my opinion, for what it's worth) turns out they do not and cannot. Other recognised formal ontological relations include identity and dependence.

Classical mereology is formal in the sense of being topic neutral; as we shall see above it is also formal in the sense of being formalised, i.e. expressed logically. The two are as Poli emphasises separable, although they are not entirely unconnected. In a perceptive discussion of the distinction he writes:

[20] This is also echoed in Albertazzi (1996), who explores Husserl's distinction between formal and material ontologies.

> Descriptive ontology concerns the collection of… *prima facie* information either in some specific domain of analysis or in general. Formal ontology distils, filters, codifies and organizes the results of descriptive ontology (in either its local or global setting)… These are pure categories that characterize aspects or types of reality and still *have nothing to do with the use of any specific formalism*. Formal codification in the strict sense is undertaken at the third level of theory construction: namely that of formalized ontology. The task here is to find the proper formal codification for the constructs descriptively acquired and formally purified.
>
> Poli 2003, p.184

The raw material for our formal ontological relations must come from somewhere, specifically from our unreflective and pre-theoretical use of particular concepts which have suitably wide application – such as part and whole. Establishing what the phenomena are to be studied is descriptive ontology. According to Poli's tripartite distinction, formal ontology operates upon this descriptive ontology to provide a theoretical framework for employing and understanding the concepts involved. The strictly optional third step is to axiomatise this framework.

We will adopt Poli's tripartite account as indicating the principal concerns of ontology: first constructing a suitable descriptive ontology for the phenomena being examined, then second analyzing this description to produce a rigorous formal theory. Rendering the formal theory in a logical format is not one of the concerns of ontology, since nothing of any substance turns on how it is done.[21] Our argument for adopting Poli's account is that it works: in exploring classical mereology we will see that the descriptive versus formal distinction is crucial to making sense of key debates over how the part-whole relation should be analysed.

[21] In what follows we will include formalised accounts of mereology and its variants, but only to provide continuity and comparison with others who do so.

2.3.1 Representation and Formal Ontology

Representation is a relation which is cross-categorial, and it is topic-neutral. There is a wealth of evidence to support this: representation takes place in the sciences, law, art, astronomy, politics, facial expressions, hand gestures, road signs etc. Even abstract objects such as numbers can be represented, e.g. as points on a manifold. Representations need not be in the same ontological category as their targets either; the passage of time can be represented with rods, for example. Furthermore, it is literally impossible to think of something which cannot be represented – the very act of thinking about that object is an act of representation itself (on either the weak or the strong reading of representation). This is true topic-neutrality, and hence representation itself could be considered to be a formal relation.[22]

Admittedly, mental representation is by definition restricted to a single domain – the mental – but this restriction makes no difference to the relation itself. Red cars and blue cars are all still cars. Identity between pigs is not fundamentally different to identity between numbers or anything else. There are different criteria of identity – for two pigs being the same has different criteria than for two numbers – but only a single relation. The same applies to representation – the criteria for one thing to represent another will vary according the things in question, but the relation is the same in all cases. For example, representing a constituency in Parliament requires being a human who is elected by public ballot while being a mental representation requires being a structured mental particular which satisfies the relations described in chapter one.

A stronger objection – made to me by Peter Simons in conversation – is that while anything can be used to represent anything else, it is arguably true that this is the case not in virtue of any intrinsic property of the representing object, but rather in virtue of it being bequeathed the status of being about something else. This, of course, trades upon the distinction between original and derived intentionality, with examples of the latter being asymmetrically dependent upon the former. The distinction is not beyond reasonable

[22] Compare with identity, the formal ontological relation *par excellence*: since identity is reflexive, everything is identical to itself. Identity is unique in that not only *can* it apply to everything, it already does.

doubt, but there do appear to be clear cases where it does apply. For example, suitable written sentences undoubtedly can represent objects and states of affairs, but there is no real prospect of identifying any intrinsic properties of the sentences or their parts underlying this fact. Furthermore, if representation genuinely were a formal relation it should at least in principle be possible to provide a formalised theory of its functioning. For twenty years a company Simons worked for sought to achieve this, without meeting unqualified success. Their failure is not proof that the task is impossible, but should certainly cast doubt on the prospects for success.

There is, however, a very compelling reason why it is necessary to engage with RTM from an ontological viewpoint. This is that RTM is itself a formal ontological theory, in exactly the sense articulated by Poli. This is best illustrated by comparison with another formal theory:

Phenomenon being analysed	*Parthood.*	*Intentionality.*
Descriptive ontology (prima facie information about the phenomenon)	Material objects are 'individuals' in a 4-D framework (see chapter three for details).	Commonsense psychology.
Formal ontology (codified and organised analysis of the descriptive ontology)	Classical mereology.	Representational Theory of Mind.
Formalised using	Predicate logic plus a primitive for parthood (see chapter three).	N/A.

Commonsense psychology, the view that intentional thought consists in sequences of propositional attitudes used to predict and explain behaviour, is a descriptive ontology: it is a collection of what is considered to be *prima facie* information about intentional thought. RTM precisely fits Poli's definition of a formal theory because it distils and codifies this

information, producing an analysis of intentionality according to which the meaning of individual thoughts is encoded in the structure of and relations between mental representations. There can be no reasonable doubt that formal ontological methods can shed light on the coherence of RTM, and that they must be used to do so.

2.4 Methods in Ontology, Again

We have seen that – to a first approximation – the principal concerns of formal ontology are formulating a sound descriptive characterisation of the phenomena being studied, and then formalising this characterisation to provide a rigorous and thorough analysis. By applying this framework to the part-whole relation in the following chapters, we will be able to both justify its value and refine it further, in order to then apply it to the concept of representation and RTM.

If it were possible, we would also apply theories of the part-whole relation to RTM and its claims about the internal structure of mental representations. Unfortunately, this is not currently an option as the precise nature of the composition required by mental representations has been almost entirely overlooked in the literature. RTMs are highly explicit about the fact that mental representations are structured and enter into compositional relations, but aside from some vague suggestions about maps on the part of Dretske and Millikan, only Fodor has anything to say about *how* representations might compose. Even this, in the form of his language of thought hypothesis, tells us little: representations need to have a syntax which is productive, systematic and compositional – as language is – but these requirements are simply too broad to be able to identify any particular type of structure or part-whole relation at play.

Using formal theories of parthood to analyse how RTMs use the concept of structure would be highly worthwhile, but it is a project which can only be undertaken once there is a significant body of data to analyse. This data does not yet exist. Our primary goal, then, is to discover and illustrate sound methodological principles in formal ontology, and then apply

them to mental representation. This will be sufficient to demonstrate that RTM is deeply methodologically flawed, and hence that mental representations do not exist.

The methods of formal ontology are difficult to establish definitively, and the practitioners of an art are not always the best sources for information on how to do the same. Rather than ask a bird how it flies, it is better to closely observe it in flight. Likewise, the best means of illustrating how formal ontology works is by demonstration. This is partly because the characterisation we will give of formal ontological method – in particular the strong emphasis we will place on empirical data – is not uncontroversial and our use of the method will serve as evidence that it is sound.

We will naturally be using parthood as the basis for our demonstration, because it is both thoroughly explored in the literature and intimately connected to representation. To structure our investigation of how to effectively explain the relation between parts and wholes, we will approach the concept using a series of problematic cases, challenging the theories examined to give a satisfactory explanation of what takes place:

1. What is the difference between a cairn and some scattered stones?
2. In what sense, if any, has a car changed if we place its engine on the back seat?
3. In what sense, if any, is half of an uncut apple *part* of the apple?

We should expect an adequate account of the part-whole relation to explain what is happening in each of these three cases. Note that these are rather mundane examples; I have deliberately avoided philosophical puzzles such as the Ship of Theseus, or the Sorites paradox, which involve parthood. The reason for this is that these puzzles involve parthood, but their solutions do not turn solely on how parthood is conceived. We might reasonably expect a theory of parthood to be helpful in explaining how to solve the Ship of Theseus, but since that puzzle is also one of identity we cannot reasonably expect a theory of parthood to dictate an answer to us.

3. Parts and Wholes: Classical Mereology

In investigating the metaphysical presuppositions of Representational Theories of Mind we have found that the central ontological claim made consistently is that they exist and are structured composite entities, i.e. that they are made up of parts. In order to fully understand how these representations are supposed to fulfil their role as surrogates for the targets of intentional thoughts, and assess its coherence, we need to have at our disposal a robust account of what it is to be a composite entity. What does it means to have parts, how do parts influence the whole(s) they compose and – in the broadest terms – how should we understand the part-whole relation or relations?

To start with, we will give a very brief historical overview of the study of parthood in the Western philosophical tradition before turning to system of 'classical mereology' which has dominated Twentieth Century debates. It is against the backdrop of this system, and our evaluation of the particular claims it makes about parts and wholes, that we will frame our conclusions regarding how the part-whole relation or relations should be understood and what methods formal ontology requires. As mentioned earlier, our approach will be themed primarily around three problem cases:

1. What is the difference between a cairn and some scattered stones?
2. In what sense, if any, has a car changed if we place its engine on the back seat?
3. In what sense, if any, is half of an uncut apple *part* of the apple?

We should expect an adequate account of parthood to answer the questions posed by the first three examples. Our discussion will not be the last word on any of these questions, or produce an account of parthood for these cases which is beyond any criticism or doubt. My aims are more modest – but ambitious enough, I think – in looking for a promising though not necessarily complete account of parthood and wholeness, one which avoids or deals with at least these problems which arise for currently dominant positions in the literature. As we will see, there are two very different analyses of parthood which can both explain all three problem cases, though for both of these analyses a number of controversial assumptions need to be made. In the remainder of this chapter we will present and explore

classical mereology, arguing that alone it cannot adequately explain any of the three problem cases.

3.1 Historical Background

The study of the relationship between parts and wholes has a long and distinguished history in Western philosophy, albeit one we shall pass over relatively quickly. Two reasons for this are firstly that our discussion will be located in contemporary discussion of parts and wholes, which with some exceptions largely ignores this historical background, and secondly that there are already a number of historical works which the interested reader can turn to.[23] Parts and wholes loom large from the days of the Pre-Socratics onwards – many or arguably all of Zeno's paradoxes involve parts, and Leucippus' and Democritus' atomism is explicitly concerned with claiming that some objects have no parts at all. Plato's *Thaetetus* and *Parmenides* as well as other works give sophisticated analyses of various instances of parthood and wholeness, and the concept is widely held to enjoy a significant place in Aristotle's *Metaphysics* (although he also makes much use of it elsewhere – by my reckoning he uses the term 'part' sixty four times in the *Nicomachean Ethics*, and one hundred and ninety five times in the *Politics*).

A catalogue of medieval to Modern philosophers who wrote on or made significant use of parts and wholes would, I suspect, read much like a list of influential thinkers *tout court* for the times, including as prominent figures Boethius, Abelard, Aquinas, Scotus, Ockham, Buridan, Spinoza, Leibniz and Kant. From what I can gather, just about every serious philosopher through these periods had something to say on the topic, and with a generally very high level of sophistication, although this fact and the details are only slowly being introduced into contemporary debates. As with intentionality, study of parts and wholes seems to have been reintroduced by Brentano and propagated by his pupils, most notably

[23] For a good survey of parts and wholes prior to the twentieth century see Burkhardt & Dufour. In particular on Plato see Harte, on Aristotle see Koslicki (2007), and for a survey of medieval and pre-modern thinkers see Arlig. For an overview of more recent developments, see e.g. Simons 2006. Some further details are present in Casati & Varzi (1994, 1999), Mann & Varzi eds. (2006), and Varzi (2006).

Husserl whose third *Logical Investigation* is generally accepted as the first real attempt to systematically study the topic.

However, Husserl's discussion is frequently taken to be lacking in rigour, and criticised for combining discussion of parts and wholes with that of ontological dependence.[24] Typically, in analytic discussions of parthood from the last few decades it is not mentioned at all – rather the starting point is almost invariably either or both of Leśniewski's *Foundations of a General Theory of Manifolds* and Leonard and Goodman's *The Calculus of Individuals*, both of which present essentially the same theory of parts and wholes in a rigorous, logical format. Grouped together under the title 'classical mereology', it is these theories which provide the backdrop for just about all contemporary discussion of parts and wholes, in particular in the last few decades through the influence of David Lewis. We will follow suit, first in presenting the theory in a format neutral between the two versions, then evaluating its principal claims. We will conclude the chapter by applying classical mereology to our three problem cases, which it will fail to explain adequately. In the chapter four we will survey the attempt to improve classical mereology by adding elements of topology, and compare this mereotopology with Kit Fine's rival analysis of parthood, the theory of embodiment.

3.2 Classical Mereology and Formal Ontology

Classical mereology and its variants have already been presented clearly and in detail by Simons (1987) and Casati & Varzi (1999) so I intend to devote only enough space to comment on a few salient aspects and outline the main tenets of the theory, before launching into discussion of the issues it raises. Classical mereology was originally motivated as a nominalistically acceptable alternative to set theory, which was perceived by the former's originators as engendering 'an orgy of double counting' (Johnston 2006, p.690). Whilst orgies are nominalistically acceptable sets aren't, so the thought was they had better not be required as the foundation for mathematics. Step in mereology: being algebraically

[24] Though see Fine (1995) for an attempt to systematise Husserl's theory more rigorously.

equivalent to set theory minus the empty set, the hope was that it could serve the same function – only minus undesirable ontological commitments, and inconvenient paradoxes.

Almost certainly its main attraction for contemporary so-called analytic philosophers is that classical mereology is presented as an axiomatised, logical theory which can be derived from a very small number of axioms. There is also no need to be a nominalist to employ it. It is both a formal theory and a formalised theory, in the senses used by Poli, and illustrates his tripartite distinction rather nicely. It is classically, and almost invariably, presented as a formalised system, but can be perfectly well understood minus symbolism. Given the formalised system's intended interpretation of modelling the parthood relation it is also formal in the sense of topic-neutral.

Mereology's descriptive ontology, unlike its formal and formalized ontologies, has received little attention. Given its nominalist motivations described above we would expect mereology's descriptive ontology to be minimal, and so it is. Everything, from the smallest part to the greatest whole, is an 'individual' – to use Leonard & Goodman's term – meaning that they are all on an ontological par. This engenders two features of the theory which are not always made immediately clear. Classical mereology is 'bottom-up' in that it is concerned with parthood rather than wholeness (or indeed parthood *and* wholeness) and this is a direct consequence of treating everything as an 'individual', combined with the principle of extensionality (see P6 below). Likewise the theory is in a sense atomistic – not in the conventional sense of taking some objects to be indivisible, which it does not in the absence of some additional axioms – but in the sense of treating each 'individual' as entirely discrete from all others, with the exception of entering into mereological relations with them.

3.2.1 Formulating Classical Mereology

Classical Mereology can be derived in numerous ways of varying length and complexity; to serve our later discussion we will cover enough but not too much, dwelling only on the key principles which have attracted the most interest in the literature. As such much will be omitted which is of interest but does not directly bear on our concerns. A considerably more detailed presentation is given by Simons (1987 ch.1-2), including a full derivation of P6 from P5 (see below), a similarly detailed account by Casati & Varzi (1999), with an overview by Varzi (2010); I see no need to waste space by reproducing large tracts from either here.

The theory is as we have said both formal and, standardly, formalised so for ease of exposition I will present the principles in both logical and non-logical format. I have no great enthusiasm for washing symbols in ontology, or philosophy more broadly, but no axe to grind about them either. Pick whichever column you prefer, or both: under the standard intended interpretation they mean the same thing. I will take P*xy* to represent '*x* is a part of *y*', assuming a standard predicate logic modulo identity, and suppressing initial universal quantifiers.

P1	*Reflexivity*	Everything is a part of itself.	P*xx*
P2	*Antisymmetry*	Parthood holds in one direction only, unless the part and whole are identical.	(P*xy* & P*yx*) → *x*=*y*
P3	*Transitivity*	If one thing is part of a second, and the second part of a third, then the first is part of the third.	(P*xy* & P*yz*) → P*xz*

Following Casati & Varzi (1999), we can call the system given by P1-P3 Ground Mereology (M). From here it is helpful to define 'proper' parthood, in contrast to the 'improper' version given by P1-P3, as being asymmetric:

$PPxy =_{df} Pxy \,\&\, \neg\, Pyx$ Proper Part

It follows that proper parthood is also irreflexive – nothing is a proper part of itself – which accords more closely with ordinary intuitions about parts. It will also be helpful in the following axioms to define overlap:

$Oxy =_{df} \exists z(Pzx \,\&\, Pzy)$ Overlap

P4	*Weak Supplementation*	If one thing is a proper part of another, then that whole has another part which does not overlap the first.	$PPxy \rightarrow \exists z(Pzy \,\&\, \neg Ozx)$
P5	*Strong Supplementation*	If an object fails to include another as a part, then the second has a part which does not overlap the first.	$\neg Pyx \rightarrow \exists z(Pzy \,\&\, \neg Ozx)$

Adding P4 to M gives Minimal Mereology (MM), which is regarded by Simons (1987, p.35) as the weakest system which can have any claim to being a theory of parts – he rules out examples such as Brentano's theory of accidents, which allows for proper parthood without any supplementary part, on the grounds that it is not entirely mereological (*ibid*., p.26). Simons' claim is debatable, even if widely agreed with, but not directly relevant to our concerns. P4 can also be derived from P5, which when combined with P1-P3 is called Extensional Mereology (EM), as it produces the following theorem:

P6	*Extensionality*	Any two objects with the same proper parts are identical.	$(\exists zPPzx \vee \exists zPPzy) \rightarrow (x{=}y \leftrightarrow \forall z(PPzx \leftrightarrow PPzy))$.

Extensionality, also sometimes called uniqueness, entails that there is no structure to mereological composition: there are no two different ways that the same parts can compose wholes, since the same (proper) parts must compose the same whole. Finally, full classical mereology (sometimes called General Extensional Mereology - GEM) is formed by

adding the unrestricted composition principle (also called the general sum or general fusion axiom):

P7	*Unrestricted Composition*	For any thing(s) which satisfy some description, there is something which also satisfies it, and which overlaps everything the original things overlap.	$\exists x Fx \rightarrow \exists x \forall y (Oyx \leftrightarrow \exists z (Fz \,\&\, Oyz))$

The same formalised system can be derived more neatly from P3, P6 and P7, and it is these three principles which have attracted the most interest and are the most controversial. As such, in this chapter we will evaluate them in detail. Incidentally, the shortest formalisation of mereology comes from Tarski (1929), with only P3 and a variant of P7 as axioms (see Casati & Varzi 1999, pp.46-7 for details), with the other principles being derivable from them. P3 and P4 were included above for convenience of exposition, and we will say no more about them. P1 and P2 are worth discussing, but turn out to be relatively straightforward. We will deal with them first, before turning to P3, P6 and P7. Lastly, classical mereology is widely associated with a further principle which is not strictly part of the theory but rather an auxiliary assumption:

P8	*Univocality*	'There is a single (fundamental) relation of parthood, which applies to all objects, regardless of ontological category' (Sider 2007, p.20 of online version).[25]

There is no compulsion to accept P8 given P1-P7, but nevertheless it is very widely accepted that they should go together, in particular by four dimensionalists such as Sider and Lewis (we will have more to say about four dimensionalism and mereology when evaluating P6 below). To indulge in a little foreshadowing, we will see that P1-P3, P6 and P7 are all defensible (though P6 requires an auxiliary assumption about the nature of material objects), but are together insufficient to tackle any of the three problem cases we began this section with. In the next chapter we will see that supplementing mereology with

[25] Sider's term for this principle is 'mereological monism'; I haven't adopted it since I find it potentially misleading, whereas 'univocality' is less ambiguous.

topological elements will go some way towards remedying this, though not far enough to constitute a full explanation of problems 1 or 2. P8 will be seen to be false.

Before turning to these five principles, a quick note on our use of the terms 'classical' and 'mereology'. Classical mereology is not remotely classical in any usual sense, being a Twentieth Century development and a major departure from theories such as Plato's and Aristotle's which might better deserve the title. However, for whatever reason the label has stuck to systems incorporating P1-P8, so we will follow that usage. 'Mereology' is ambiguous between two meanings: first the study of parts and wholes broadly conceived, and second more narrowly systems to which classical mereology is closely related. We will stick to the second meaning, with 'mereology' proper restricted to GEM (P1-P7) and any stronger system which incorporates it. This should help avoid confusion; for example, van Inwagen (2006) argues that 'mereological sums' may change their parts, despite that being apparently ruled out on technical grounds (see Casati & Varzi 1999 p.46 for details). In our terms, the summation principle he is concerned with is non-mereological, although it would be mereological in the first, broader, sense of being concerned with composition.

3.3 Reflexivity and Antisymmetry

> Many legitimate senses of 'part' are nonreflexive, and do not countenance saying that a whole is a part (in the sense in question) of itself. The biologists' use of 'part' for the functional subunits of an organism is a case in point.
>
> Rescher 1955, p.10

This problem is generally dismissed as trivial, as 'mere verbal quibbling' (Simons 1987, p.107). Since an irreflexive 'proper' parthood relation can be readily defined as 'improper' parthood without identity, the issue becomes that of which primitive concept to choose (see e.g. Casati & Varzi 1999, Simons 1987, Varzi 2010).[26] The same formal system, i.e. ones which share all their well-formed formulae, can be produced from a range of different

[26] If 'x is a (reflexive) improper part of y' is represented by 'x < y' then we can define 'x is an (irreflexive) proper part' as 'x < y & ¬(y < x)'.

primitives including proper or improper parthood as well as less immediately obvious candidates such as overlap. The various options are surveyed nicely in Simons (1987) ch.2.

There are two reasons worth mentioning why the choice between primitives is not entirely trivial, though neither is terribly substantive. First, varying interpretations of a single concept such as parthood allow the possibility of equivocation and confusion – we need to take care to avoid either happening. This is of course achievable with care; it may be that many or most uses of 'part' outside of mereology are irreflexive, but provided we can translate their claims in terms of a reflexive concept *salva veritate* then there should be no problem provided it is done explicitly. For this pragmatic reason, though phrased in different terms, Johansson (2006) suggests using proper parthood as a primitive, and I am inclined to agree it would make life easier. To be consistent with general usage in the literature, however, unless explicitly said otherwise we will be treating parthood as improper.

The second reason is a little more significant, being that the definition of proper parthood will – if transitive – necessarily be asymmetric. If there are legitimate uses of 'part' which are reflexive and symmetric then this will not do for those cases, and so a mereology based upon proper parthood or any primitive interdefinable with it will lose universality – it won't apply here. This is, however, just the question of whether mereology is right to take parthood to be symmetric.

It seems unlikely that there are any real cases of symmetric parthood. At least, I know of none which have been seriously suggested, and only brief reflection is needed to see that ordinary usage seems not to allow for it. But what about extraordinary usage? Sanford (1993, p.222) cites Borges' Aleph: 'I saw the earth in the Aleph and in the earth the Aleph once more and the earth in the Aleph…' (Borges 1949, p.151, quoted in Sanford 1993). The 'saw in' relation used here is symmetric, and assuming it is interpreted as parthood it presents an apparent counterexample to asymmetry.

That assumption may be too strong though. I saw my reflection in the mirror this morning, but it doesn't follow that the reflection is *part of* the mirror. For one thing, it isn't present very often (only on the rare occasions when I am too), and when it is it arguably tends to

appear to be located a foot or two behind the mirror. Furthermore, even if the Aleph is accepted as a description of symmetric parthood, Casati & Varzi (1999, pp.35-6) and Varzi (2010) point out that there is no compulsion to treat literary fiction as an accurate guide to metaphysical possibility. To use a little jargon, even if conceivability is a guide to possibility, grammatical and lexical soundness is not a guide to conceivability (a point which, to digress, seems to be somewhat lost on David Chalmers).

This seems a highly reasonable claim: lay (by which I mean non-mereological) uses of 'part' are certainly nonsymmetric. For current purposes, then, we are safe enough to assume that (improper) parthood is antisymmetric, and so that P1 and P2 stand.

3.4 Is Parthood Transitive?

Whether parthood is a transitive relation is rather more controversial than whether it is reflexive or antisymmetric, owing to the existence of numerous problematic examples which appear to show it is not always so. In this section I will argue that there is a strategy which successfully deals with all such examples whilst maintaining the transitivity of parthood. To show this we will consider four problematic examples. The first comes from Rescher, the second Cruse, and the third and fourth from Johansson (2004). There are numerous others; Johansson gives a detailed discussion including a further ten, but our selection combine the most well-discussed cases (a & b) with what he rightly considers to be the most problematic (c & d). The 'standard' approach discussed below deals successfully with all four cases.

Quite a few examples have appeared in the literature, so an exhaustive survey would be overly long. We shall consider the four most challenging ones:

a) A nucleus is part of a cell, and the cell is part of an organ, but the nucleus is not part of an organ.
b) A handle may be part of a door, and the door part of a house, but nevertheless the handle is not part of the house.

c) X may be a large spatial part of y, and y a large spatial part of z, but x need not be a large spatial part of z.[27]
d) If x is a spatial 60%-part of y, and y is a spatial 60%-part of z, then x cannot be a spatial 60%-part of z.

The challenge presented by a) – d) is quite straightforward. If mereology accurately models a formal ontological relation of parthood, then its axioms must be true of *every* part of *every* whole. If it really is true that, for example, a handle may be part of a door and the door part of a house, but the handle not part of the house, then this seemingly presents a counterexample to transitivity as a universal feature of all parthood instances, and so any mereology which includes it as an axiom.

Taken at face value, these examples and others like them seem to require abandoning either or both of the features of mereology which render them problematic in the first place: transitivity (P3) and univocality (P8). As such, the part-whole relation would either be non-transitive in general, perhaps with some particular instances being transitive or intransitive, or there would need to be multiple part-whole relations. In the latter case, it might be that each relation is restricted to certain kinds of entities and is transitive (Winston, Chaffin & Herrmann), or some relations may be transitive and others not. Other approaches are no doubt possible as well.

[27] A quick note on c) and d): 'spatial part' is a somewhat ambiguous term, between either a part of the space occupied by an object, or an object which occupies part of the space occupied by an object (or, I suppose, both on a suitable conception of space). For present purposes it makes no difference which interpretation we choose.

3.4.1 The 'Standard' Account

There is however a response to these problem examples, suggested by Simons (1987) and endorsed by Casati & Varzi (1999) and Varzi (2010), which seeks to rationalise the mereological claim that there is a single parthood relation which is transitive with acceptance that some instances of parthood are not.[28] Following Varzi (2010) I shall call this the standard approach. This takes the parthood relation to be univocal, but recognises that many statements about when some things are parts of others to be relative to a particular set of interests, or in Casati & Varzi's terminology to 'count policies'. Reflecting these interests, some sort of restriction φ is placed upon the parthood relation, to give a new narrower relation of φ-parthood. The narrower relation could be transitive, intransitive or non-transitive, depending on the meaning of φ. In contrast, the broad and unrestricted parthood relation is always transitive. For example, take φ to be 'at the outside border of'. A brick may be at the outside border of a wall it is part of, and a bit of clay at the outside edge of the brick it is part of, but the bit of clay is not necessarily at the outside border of the wall. It might be next to some mortar. Here parthood is transitive, but the 'at the outside border of' relation is not, and so their combination is non-transitive.

But just what sort of combination of relations is taking place here? An easy interpretation of φ-parthood is that their combination is akin to the relationship between a determinable and a determinate, such as having colour on the one hand and being green on the other. If that were the case, it would be difficult to see how transitivity would not remain true (be 'passed down' in Varzi's terminology) regardless of the restriction. A ball is still a ball, whatever colour it might be, so shouldn't a φ-part still be a part, whatever φ might be? And mereology tells us that parthood is a transitive relation. On this interpretation, which was pressed against the standard account by Johansson (2004) then rescinded in his (2006), the standard account just doesn't work.

Careful reading of Casati & Varzi (1999) reveals a different interpretation which is much more tenable: in our theorising about parts and wholes we should adopt a broadly Quinean

[28] See also Lewis (1986) p.213 for a similar strategy in the context of restricting composition.

view of existence, as being whatever is included in our domain of quantification (see Varzi 2000 p.5 of online version for a clear statement of this point). Different interests, such as physiology versus astronomy, will be concerned with different domains, and as such our theories of part-whole relations will in each case range over different entities. So the restriction in the standard account is a domain restriction – parthood in the classical mereological sense has an open domain, which is narrowed in the case of φ-parthood to only include those things which are φ. Some properties will be unchanged (or will 'pass down') through this restriction, but plausibly enough transitivity in many cases is not one of them (see Varzi 2006 §2).

This is a somewhat technical presentation, so if it is not entirely clear try this version. Imagine a really big Venn diagram which incorporates absolutely *every thing* (note the space between 'every' and 'thing'). Lots of circles represent different interests, some of which overlap or underlap others. If you are a biologist, for example, the only parts you care about are the ones in the biology circle, which are both parts and biological in nature (this is the φ-restriction). Inside this circle may be another circle for marine biologists, and here every thing is first of all a part *simpliciter*, but also a biological one and a marine one as well (this is a more stringent φ-restriction). Much of what is true of parts in general is true of marine biological parts, e.g. that they enter mereological relations, and much of what is true of biological parts in general is true of marine biological parts as well – e.g. at some levels of decomposition they are cellular. However, not everything which is true of marine biological parts is true of other parts under different (or no) restrictions. Transitivity may not be true of marine biological parts even though it is true of parthood *simpliciter*: a lepton is a part of a dolphin which is part of a school, but a lepton is not a *marine biological* part of a dolphin. When we look at the whole Venn diagram, we can see that parthood across all circles is transitive, but when we narrow our attention to the contents of particular circles it may well appear to be intransitive or non-transitive, because those circles leaves lots of things out. I don't know how to make the point any clearer than this, so let's move on and apply the schema to examples a) and b).

3.4.2 The Standard Account: Example a

a) A nucleus is part of a cell, and the cell is part of an organ, but the nucleus is not part of an organ.

According to the standard account, it must be plausible to interpret example a) as involving an unrestricted parthood relation which is transitive, and I shall take it that this is correct. It would be unreasonable to deny that such an interpretation is available at all, and the standard account does not entail that the classical mereological sense of part is one which is used frequently in ordinary contexts, if indeed at all outside of discussions like this one.

To apply the standard account to a), then, what we need to do is supply some sort of domain restriction which includes nuclei, cells and organs, and a relational property satisfied uniquely by at least some of our new domain's members. Preferably the resulting analysis should accord reasonably well with some intuitive reading of the example – without this restriction, we could just cook up a restricted domain and some gerrymandered property to go with it, which would be technically sufficient but quite unhelpful. I suppose a superficially plausible candidate for an intuitively acceptable domain restriction is that of spatial entities, with a relational property of *x* fully overlapping *y*. Organs, cells and nuclei occupy space and the larger fully overlap the smaller ones. The only problem with this is that it fails to account for why a) seems problematic, as in this case nuclei presumably would turn out to be parts of organs. What we want is an account which recognises why a) looks like a case of non-transitive parthood, but which tells us why it actually isn't.

For a) to look like an example of a non-transitive parthood relation, what is needed is an understanding of the relation involving some sort of function or organisation. A cell isn't just part of an organ, it's a part which contributes directly to the organ's working. Likewise for a nucleus and a cell. However, a nucleus plausibly does not directly contribute to the working of an organ it is within, and this is why the example appears to contradict the transitivity of parthood. A satisfactory analysis following the standard account needs to incorporate this in some way. This could be done by restricting the domain to parts of organisms (with 'parts' suitably cashed out here to avoid a regress), and choosing a relational property such as '*x*

contributes to the functioning of *y*'. This is admittedly a rather vague property, but plausibly it can be cashed out in a way which makes the resulting restricted parthood relation intransitive. Parthood simpliciter remains transitive, so the standard account succeeds in dealing with example a).

I do think there is a little more to be said about the example though, in the name of a more detailed analysis. With only a little effort we can give a better analysis which is still compatible with the standard approach, by observing a significant difference between the ways in which 'part' is applied in the example to nuclei and cells, and to cells and organs. Compare a dialysis machine to a liver – it may be too much of a stretch to say the former is literally an organ, even though it replicates (at least some of) the functions of the latter. But a sufficiently small dialysis machine implanted in one's body (and, perhaps, perfected to replicate the liver's function exactly) seems like a pretty strong candidate for organhood. I see no good reason to legislate that organs must be organic, hence made of cells or things analogous to them, despite the common etymology. So an organ doesn't necessarily need to have cells as its parts; it just needs to work.

On the other hand, the case for a nucleus being an essential part of a cell seems to me to be much stronger. The nucleus could be synthetic, perhaps made of inorganic elements if that's physically possible, but without a nucleus a cell simply couldn't function properly *as a cell* – it couldn't engage in mitosis or meiosis for starters. This highlights a definite equivocation in the first two uses of 'part' in example a); given this, it is not surprising that trying to read all three uses as the same unequivocal relation is awkward. So a better analysis simply requires pointing out this equivocation, and doing so is fully compatible with the standard approach. The trick is to realise the equivocation implies two distinct domain restrictions on the same relational property (along the lines of '*x* contributes to the functioning of *y*'). Based on the dialysis machine argument just mooted, I think it likely that restrictions of 'essential parts of organisms' and 'non-essential parts of organisms' respectively would do the job (again, with the proviso that 'parts' in each of these is suitably cashed out). Treating these two restrictions as equivalent, i.e. thinking the uses of 'part' in example a) are univocal, produces the apparent intransitivity of parthood in the example. Recognising the equivocation over

'part' allows that both relations may be transitive in their own domains, but intransitive in combination.[29] Simple!

3.4.3 The Standard Account: Example b

b) A handle may be part of a door, and the door part of a house, but nevertheless the handle is not part of the house.

To indulge in an anecdote, on returning from honeymoon my wife and I found that someone had made an abortive attempt to burgle our house, which involved removing and making away with the front door handle. I am quite sure that a part of our house was taken, because without the door handle it proved very difficult to get inside. If the house hadn't lost any part of itself (and not undergone any other relevant changes, which it hadn't), it shouldn't have been any different – and different it certainly was. This accords with a transitive analysis of parthood in b), although we should bear in mind that the standard account claims not just that there is *some* applicable sense of part which is transitive, but that it also satisfies the other axioms of classical mereology. I take it that this particular interpretation does. The honeymoon case shows it is perfectly reasonable to take the parthood relations in example b) to be transitive, and the statement made in b) to be false.

That is not to say that houses necessarily need door handles, nor that where there are handles present they must be a part of the house. Rapunzel's tower didn't have a door handle, or a door for that matter, but it served as a house well enough. Some houses may use only sliding or rotating doors, and to insist that igloos are not houses would be spurious and overly revisionary of the ordinary concept. Houses do not need doors, we should say, but where they have (working) doors the doors are a part of them. The working qualification is significant, as it captures the intuition that simply being a door inside a house is not enough – the door needs to fill a suitably shaped hole in a wall and be capable of being

[29] To head off one possible complaint about this argument, I'm not saying that a relation of essential parthood definitely is transitive, just that if it were then when run together with a different transitive parthood relation the result might be intransitive. Essential parthood could be non-transitive or intransitive and the same would also be true.

opened and closed. Likewise a door need not have a handle, but when it has got one which works the handle is a part of the door.

Example b), then, can be reasonably analysed using the standard account: as employing a restricted notion of parthood, that of a working (or functional) part. We can apply Varzi's schema to it, with F*xy* representing a relational property such as '*x* contributes to the functioning of *y*', although we would have to cash this out a little differently than for a) (which is no great surprise, since the parts in a) are organic rather than mechanical). I am inclined to think that again this restricted parthood relation is in fact transitive: whenever a house has a working door which has a working handle, the handle is a working part of the house because it contributes to the door's function of letting people in the house, which contributes to the house's function of being suitable to be lived in. It's tricky living in a house which needs a door handle but doesn't have one – I should know. For a house without any doors, or with doors which don't need handles to function, there is no handle to be a relata, so no parthood relation to argue over. This view can be happily accommodated by the standard account by taking the 'working' restriction to be itself transitive, as likewise can the view that a) is a genuine case of intransitivity. Whether 'working part' is transitive or otherwise is a question for the theory of functions, something I won't dwell upon further here except to say I do not rule out intransitive or nontransitive functionality in other cases.[30]

3.5 Johansson on the Arity of Parthood

We have seen that the standard account deals quite well with the standard problem cases a) and b), but more recently Johansson (2004) has proposed a novel account along with two new and more 'conspicuous' (2006) examples – our c) and d) – which he takes to favour his own account over the standard one. The core of Johansson's position is that parthood is a binary relation when transitive but, contra Varzi and others, intransitive and nontransitive

[30] For some further discussion of functions and functional parthood see chapter four.

parthood relations are ternary.[31] That is to say that while the parthood relation looks binary as it only makes explicit reference to a part and a whole, in intransitive and nontransitive cases there is implicit reference to a third object over and above those two, which must be present to 'ground' or make possible the parthood relation itself.

Varzi (2006) does in fact agree that at least some parthood relations involve a "hidden and indefinite" (p.7) reference to a third object; for example in b) the handle does not act upon the whole door but only on a panel of it. Nevertheless he maintains that parthood is always a binary relation. To see why, recall the Quinean view of what exists being given by the domain of our theory; according to Varzi the implicit third object is bound by an existential quantifier whereas the door and handle are not, hence the relation is binary. In response Johansson (2006) has, I think rightly, pointed out that this distinction is terminological: we need only replace variables with constants for a particular door, handle and panel to produce an arity of three. The substantive issue is whether there are any intransitive or nontransitive φ-parthoods which do not on analysis "reveal some kind of relative product" (p.3 of online version).

Given that the standard account can deal with cases such as a) and b) easily enough, the crux of Johansson's argument for his position turns on his discussion of spatial parts through the examples of c) and d):

> To be large is to be large in comparison with something. Therefore, there must in the case at hand be an entity, z, distinct from x and y, in relation to which x is large.
>
> Johansson 2006, p.3

Here *z* is the implicit third relata which makes the (intransitive or nontransitive) parthood relation ternary. Johansson's first sentence is a truism, but I dispute that it entails the second. It does not follow in general that the something *x* is large in comparison to must be distinct from *x* and *y*, nor does it follow in the specific examples of c) and d). In fact, we shall see that if we assume parthood to be a relative product or a ternary relation, c) collapses

[31] Ternary parthood relations are originally compared to qualified relative products (2004), but later also to regular relative products and primitively ternary relations (2006). We don't need to worry about the details here – the important issue is that parthood is taken to involve three relata (2006, p.3 of online version)

into d) on pain of circularity or regress, while d) simply does not require the existence of a third 'comparison' object. Johansson's ternary account fails to deal with his own examples, while the standard account succeeds.

3.5.1 Large Spatial Parts

c) x may be a large spatial part of y, and y a large spatial part of z, but x need not be a large spatial part of z

Johansson argues that by analogy with '*a* is an aunt of *b*', 'large spatial part of' in c) should be considered to be a relative product – just as in order to have an aunt one must have or have had a parent, "necessarily there is at least one object of size comparison (C*w*) such that '*a* is larger than *w*' (*a*L*w*)' (2004 p.6). Furthermore, *w* must be distinct from *a* and the further object they are both parts of (2006). But there is a striking disanalogy here. Having a parent is a 'grounded' relation, one which does not require any more than that there be a child and have been a parent. By extension we may say that the relationship between child and aunt is grounded as well – by the existence of both, and their relationship with the parent. The relationship between a large spatial part and the third 'comparison' object, however, is not necessarily grounded in this sense. Johansson imposes no restrictions on what it might be – it just needs to be relatively large. This opens up the prospect of any bit of detritus being press-ganged into a ternary parthood relationship. What has Gandhi's toe bone got to do with the parthood relationship between a liver and a large spatial part of it? Nothing at all, except that our part of a liver is larger than Gandhi's toe bone – and for Johansson's account that is enough. Worse than this arbitrariness, however, is indeterminacy: the fact is that innumerable objects would do the job just as well and it isn't clear which the right one is, or even if there is a right one to be had. Formal and formalised theories are supposed to help sharpen up and clarify our sometimes unclear and confusing thoughts – vague or indeterminate relata are definitely to be avoided.

There seems to be a happy work-around to both of these worries, though. Why not take the comparison object to be another part of the whole – another spatial part since we're comparing size? That way, the 'grounding' problem is circumvented as the third object is just as 'grounded' as the other part. It is also certainly intimately related to the original two relata for the parthood relation, dispelling the threat of arbitrariness. Admittedly, it might still be a bit vague exactly *which* other part we are using for comparison, but perhaps we can let that worry go as there are a smaller number of reasonable candidates.

However, more problems arise. As we have observed, the third object must presumably be itself a spatial part of the whole to allow for size comparison, and it is going to be a large, medium or small part (or very large, very small etc…). If it is a large spatial part, then we are presented with no analysis at all of the relation in c), or at least one which is circular: "to be a large spatial part is to be related to another large spatial part". This is highly uninformative. If the third object is a medium or small spatial part, it seems Johansson's account will classify them as relative products as well. But then for something *x* to be a large spatial part of *y* is for there to be *z*, a smaller spatial part of *y*. And for there to be something *z* which is a smaller spatial part of *y*, there must be yet another thing which a smaller part again. Unless we make the unreasonable assumption that all spatial objects have a smallest part, this clearly engenders a regress.[32] Worse, it is a distinctly vicious one. That x's *being a large spatial part* is dependent upon the existence of a smaller part is not automatically problematic, but I take it that when *every one* of the dependency relations between our many spatial parts requires some further part, then the whole series of relations is itself unsupported and the regress is vicious.[33]

Analysing 'large spatial part' as a kind of relative product or ternary relation then is either circular or viciously regressive, at least on a qualitative analysis using terms like 'large' 'medium' and 'small'. An alternative is to cash out these terms quantitatively, and this

[32] Jonathan Tallant has pointed out to me that this could in fact be used as an argument in favour of all objects having smallest parts. I suppose it could, but to do so looks suspiciously like a tail wagging the dog – we would be starting out with a quite specific problematic example, and ending up making a very broad claim about parthood in general.

[33] This account of vicious circularity is adapted from Anna- Sofia Maurin's 'Infinite Regress: Virtue or Vice?' presented at the 'Methodological Issues in Contemporary Metaphysics' conference, Nottingham, 7th January 2006. Maurin's account differs in that it is limited to existential rather than relative dependence, that is on an object depending for its very existence rather than some of its properties upon another object.

would provide us with an analysis along the lines of '*x* is a large spatial part of *y* in virtue of being bigger than *z*, which is a spatial 60%-part of *y*.' Thus on Johansson's account, either the parthood relation in c) is ungrounded, or to all intents and purposes the example collapses into d).

3.5.2 60%-Spatial Parts

d) If x is a spatial 60%-part of y, and y is a spatial 60%-part of z, then x cannot be a spatial 60%-part of z.

Johansson (2004) comments 'I guess and hope that no further arguments are now needed to show that, just like the predicate 'large spatial part of', the predicate 'spatial 60%-part of' designates a relative product' (p.7). Since his analysis either fails outright for 'large spatial part' or else runs the two together, further arguments are very much needed. However, I very much doubt that any sound ones will be forthcoming.

Imagine a bag containing ten balls and nothing else – if six balls are black then 60% of the contents of the bag are black. It is surely false to suggest that this is true in virtue of some other object – some other collection of balls, whether in the same bag or not. The claim is true in virtue of there being ten balls in the bag, six of which are black. Presumably, Johansson would be happy to agree with us here. However, the same principle plausibly carries over to the space occupied by the six balls – 60% of the space occupied by balls in the bag is occupied by black ones in virtue of there being ten (equally sized) balls in the bag, six of which are black. I cannot see how there is any need for – or even any sense in – positing additional objects to explain this straightforward fact.

But if 60% of the space occupied by balls is occupied by black balls in virtue of there being ten balls in the bag, six of which are black, then surely the fact that the total space occupied by black balls is 60% of the space occupied by all the balls is true in virtue of just the same fact. To express this more succinctly, the black balls occupy a spatial 60%-part of the

contents of the bag.[34] I see no need to introduce any entity over and above the six black balls and the ten balls in total, nor any advantage to doing so (the bag doesn't count, as it is just a way of limiting our domain – we could just as well imagine a possible world containing ten balls[35]).

3.6 The Standard Account Again

Johansson's own ternary account of parthood fails to account for his own problem examples. To the best of my knowledge there has been no attempt to apply the standard account to c) or d), with Johansson laying down a challenge to do so:

> I would very much like to see an analysis of my most conspicuous example of non-transitive parthood, 'x is a large spatial part of y', that does not (just like 'x is a grandparent of y') reveal some kind of relative product.
>
> (2006, p.3)

In the following section I shall apply the standard account to c) and d) to give an analysis of their parthood relations which is strictly binary without being transitive. I shall begin with d); if the above arguments stand then this is the case Johansson should be more interested in anyway.

Let's think back to our Venn diagram analogy. We can think of the whole diagram incorporating parts simpliciter with individual circles for various restrictions. One quite large circle would represent spatial parts, with all non-spatial parts on the outside. Parthood simpliciter – i.e. ignoring all circles – is transitive, and plausibly enough spatial parthood is as well when there are no further restrictions in place. If 60%-spatial parthood were represented as a further circle, and I think this is how Johansson views the standard account, then it would indeed fail to demonstrate non-transitivity. But this isn't right.

[34] To use the alternative interpretation of 'spatial part' mentioned earlier, we could instead say that the space occupied by the black balls is a spatial 60%-part of the space occupied by all the balls.

[35] I'm choosing to ignore the possibility that substantival space-time could be considered an object itself, on the grounds that while this is a possibility it is only one of many ways of conceiving space and time, and one which needs explicit motivation. Thanks to Jonathan Tallant for bringing this point to my attention.

'Spatial part' describes a property which is the same for everything it applies to: each is a part which is spatial. The additional restriction of 60%- produces a relational property which is not the same in different cases; this much should be obvious when we consider that 60% of the matter in the universe is a very different sort of quantity to 60% of the matter in my stomach, both in terms of possible and actual magnitude, to say nothing of variation over time. Φ-restrictions upon parthood are not all equal, there are different types according to the sorts of properties or relations which are employed as restrictions. Relational properties such as 60%- are going to generate non-transitive or intransitive φ-parthood relations. The relational nature of 60%-spatial parthood shows a germ of truth in Johansson's account, but *contra* Johansson there are only two relata.

The same process applies mutatis mutandis to 'large spatial part' in c). Large spatial parts are parts which are spatial and which are large. Non-transitivity comes from the restriction of largeness being a relational property, and again none of the relations call for a third object. So the standard account can likewise be readily applied to c).

3.6 (Unrestricted) Parthood is Transitive

By using the standard account of Varzi and others, each of the problem examples can be analysed effectively to be either intransitive or nontransitive without contradicting the transitivity of the formal mereological parthood relation. This is because the mereological relation is conceived as taking an entirely open domain, which is restricted in one of two ways: either by monadic properties reflecting particular areas of interest, or by relational properties which hold between some combinations of parts and wholes but not others. Apparent cases of intransitivity or non-transitivity can arise through three different circumstances. Firstly, through an implicit move between different monadic restrictions such that there is an equivocation between uses of 'part', such as in example a). Secondly, through a questionable interpretation of a specific case which is actually transitive, as in example b). Thirdly, through a relational restriction being misinterpreted, as in c) and d). I cannot take any credit for originality here; as I see it the key ideas expressed in this section

have already been put forward by Varzi (2006). All I have done is presented and explained them somewhat differently, making use of two additional examples. My diagnosis of the four cases does differ in that, if I understand him correctly, Varzi treats a) – d), and indeed most cases, as examples of restrictions using relational properties, with this being the source of their intransitivity or non-transitivity. Either way, the transitivity of mereological parthood simpliciter stands.

3.8 Mereological Extensionality: Is Parthood Structured?

Is mereological extensionality true or false? Another way of approaching the same issue is to ask how many non-identical objects can be made, or 'composed', out of the same parts? The answer given by classical mereology is that there is only one; extensionality tells us that any two objects which share all and only the same parts are in fact identical. This entails that there is no structure or organization to the way composite objects are composed from their parts, as there are no two different ways that things may come together to form a whole. Readers may detect some ambiguity in the use of 'same' and 'different' with regards to parthood, however – we should be clear that axiomatic extensional mereology is committed only to the claim that numerical identity of parts entails numerical identity of the wholes they compose. But taking on extensionality as a philosophical thesis it would be odd to say that while this strict numerical claim is true the qualitative equivalent is false – that is that two strictly non-identical wholes with qualitatively identical parts may themselves be qualitatively non-identical. Subsequent discussion of extensionality will address the qualitative extensionality principle, as well as the numerical one.

Whatever the status of nominalism in contemporary philosophy, extensional mereology has endured as a popular account of the relationship between parts and wholes, either in its classical form (see e.g. Lewis) or under some variation which preserves the features mentioned (such as with topological elements – see chapter four), and given this popularity a significant critical literature has grown. Within the literature numerous intended counterexamples to (strict numerical) extensionality have been proposed, but they have

failed to achieve widespread acceptance. I think this is entirely appropriate, because no example could even *in principle* disprove extensionality. To show why this is the case, I will survey some of the existing intended counterexamples and canvass the *prima facie* plausible suggestion that they fail on the grounds of questionable ontological commitment.

Drawing on molecular chemistry I will propose and motivate a novel example parallel to those in the literature, but of less dubious ontological status, before showing that even this stronger case ultimately fails to provide a counterexample to either strict or qualitative extensionality. Weak arguments are not always terribly interesting in themselves, but considering why the novel case fails will show us that extensionality simply cannot be disproved by example, and so all attempts to do so are entirely futile. Furthermore, it will be seen that whether we accept or reject extensionality depends upon how we characterize the composite objects we are considering, and consequently the choice is not substantive. This fact illustrates an important point about method in ontology: one's choice of how to characterise material objects has the capacity to shape future debates over those objects' properties, and can even determine what theoretical options are available.

Since parthood is a formal relation, one which applies to every category of being - or, in less provocative terms, one which is topic-neutral - it is not surprising that putative counterexamples to extensionality have been suggested invoking a wide range of objects, including cats, statues, orchestras, and sentences.[36] What they do have in common, however, is the form 'objects *A* and *B* share all and only the same parts but are nevertheless not identical, as they bear different properties'.[37] We can substitute various kinds of entities for *A* and *B*, and different sorts of properties, but the basic format remains the same. Thus, particular symphonic and wind orchestras may in fact share all the same members, yet serve

[36] Some philosophers restrict the scope of extensional mereology, most commonly to quantities of matter (e.g. Needham, Simons). The intended counterexamples described here only address extensionality either as a universal claim about all instances of parthood, or restricted to the domains of those examples, so have no impact on this position. They also have no impact on the view that extensional mereology is universal but is one of numerous part-whole relations, which I favour but will not argue for here.

[37] Admittedly, Rescher's intended counterexample of the sentences 'John loves Mary' and 'Mary loves John' may not fit as neatly into this characterisation, but it is widely regarded as failing since the words in each sentence may be regarded as distinct tokens of the same three types. Furthermore, the sentences plausibly do not have all the same parts: 'John loves', for example, features in one sentence and not the other.

different functions (perhaps they perform distinct repertoires on separate occasions), or a given statue may be composed of the same matter as a given lump of clay, but the two nevertheless have quite different persistence conditions.[38] I won't rehearse the reasoning behind these sorts of cases, but for discussion of the former see in particular Simons (1987), and for the latter see e.g. Baker, Lowe (1989), Sanford (2003) and Wiggins.

So why have these intended counterexamples failed to command universal, or even particularly widespread, assent? An initially plausible suggestion comes from observing that advocacy of extensional mereology often goes hand in hand with some form of ontological austerity, such as nominalism. In the light of such prior commitment the intended counterexamples lack force. One might, for example, believe there are principled reasons to doubt the existence of orchestras – at least as being anything over and above their members – on the grounds, perhaps, that composition is in fact identity, or bears a very close analogy to identity (see in particular Lewis 1991). Somewhat similarly one might believe that there are no such things as statues, or even things like hands (for defence of these views see van Inwagen 1990 and Olsen respectively). Thus we are presented with what looks very much like a clash of intuitions, of starting positions – the intended counterexamples to mereological extensionality may well be persuasive to a philosopher who is already committed to the sorts of entities which figure in them, but not at all to others who lack these commitments – thereby generating an impasse. Following this line of reasoning, the trick to bridging this divide, so arguing more persuasively against mereological extensionality, would be to produce a parallel example – one with the same form – involving entities of relatively uncontroversial ontological status. Just such a case can be drawn from the chemical phenomenon of optical isomerism.[39]

I will construct a best case argument for optical isomerism presenting a genuine counterexample, then discuss why the argument nevertheless fails – and what this tells us about the dispute over extensionality. In brief, the argument fails because optical isomerism can be made to be compatible or incompatible with extensionality depending on prior

[38] I ignore here the possibility that membership of an organisation or group may be a distinct relation to parthood.

[39] In fact comparable examples can also be drawn from other forms of isomerism, but optical isomerism is adopted here as it presents a particularly clear case.

assumptions about the nature – *not the existence* – of composite objects. The disagreement turns not on the ontological question of what there is, but of what it's like, with the existence of the 'it's in question being taken for granted. Another way of making the same point would be to say that, despite appearances, the impasse over extensional mereology is generated by disagreement over commitment to certain types of properties, rather than to any particular entities. But that remains to be shown, so back to our new intended counterexample.

3.9 Optical Isomerism as a Counterexample

Optical isomerism is a structural phenomenon which can occur in certain types of molecules – typically including at least one carbon atom – such that its atoms may be arranged in either of two configurations, with one being the mirror image of the other. Thus in figure 1 below (A) and (B) are optical isomers of each other:

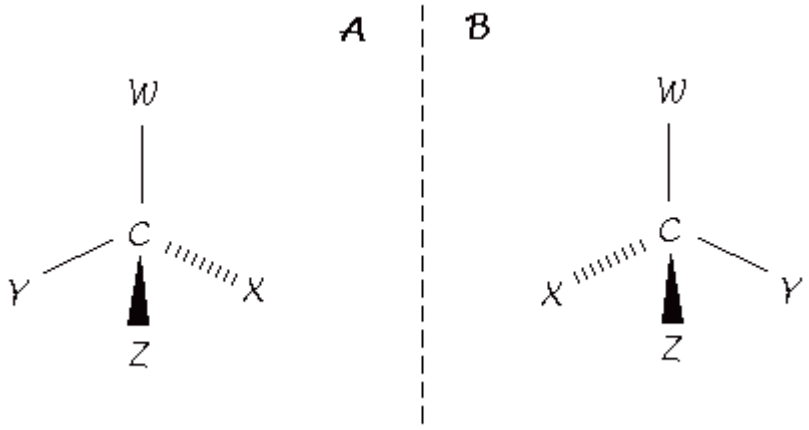

Figure 1: An illustration of optical isomerism.

The four bonds connecting *W*, *X*, *Y* and *Z* to the carbon atom *C* are (roughly) equidistant in three dimensions, with the four letters each representing different groups of one or more atoms. As with our left and right hands, neither molecule can be superimposed on its mirror image. Nor can they be rearranged by any process – such as rotation – which does not involve disassembly and reassembly. Optical isomerism is relatively common in organic

chemistry, being exploited by many living organisms. For example, the quite different flavours of orange and peppermint oil are produced by optical isomers of the hydrocarbon limonene ($C_{10}H_{16}$). Both isomers are found in turpentine, which – I suspect – tastes quite unlike either orange or peppermint. Likewise caraway and spearmint are products of optical isomerism. A better known, and more serious, case is that of thalidomide ($C_{13}H_{10}N_2O_4$), a drug which in the form of one optical isomer produces an anti-emetic effect, whereas the other isomer is a teratogen – it produces severe malformations in unborn infants when ingested by pregnant women.

The phenomenon of optical isomerism differs from the intended counterexamples above in that it makes use of entities which feature in all but the most austere of ontologies, but may be used to construct an example of the same form, i.e. 'objects *A* and *B* share all and only the same parts but are nevertheless not identical, as they bear different properties'. The argument, then, would run something like this:

Premise 1: Optical isomers differ in their physical properties.
Premise 2: They share all and only the same parts.
Conclusion: Extensionality is false in this case, and so as a universal principle.

There are, I think, good reasons to think that both premises are true, although the first turns out as being easier to motivate than the second. We will consider them in turn, before questioning whether this is enough to justify belief in the conclusion. Before we do so, there is another, independent, moral to be drawn from optical isomerism. The difference between left-handed and right-handed isomers is purely geometrical, and so cannot be accounted for by mereology alone. As Peter Simons has pointed out to me (in conversation), this point generalises and is not restricted to isomerism: in all dimensions above zero there is a binary distinction between two modes of orientation. Consequently, there must be spatial categories – and hence not just mereology – in any formal ontology which includes objects existing in one or more dimensions. We will return to how spacial categories could be used to complement mereology in chapter four.

3.9.1 Premise One

So why must optical isomers have differing physical properties? Consider how in general we tell the difference between two molecules: the most obvious way is by their chemical formulae. Hence the difference in composition between carbon dioxide and sodium chloride is represented by their respective formulae of CO_2 and NaCl. But by definition isomers share the same formula, so that won't work – unless we include in the formulae some indication of which isomer they refer to, but that would only describe the difference, not justify claiming that there is one. As a more promising option, we might think that the various properties of volumes of chemical substances arise from differences in the properties of the individual molecules from which they are composed. Hence differences in solutions of a particular isomer, such as tasting of orange or peppermint, might depend upon the properties of the individual isomers present. All optical isomers can be individuated as either type (+) or (-) according to whether they rotate plane polarised light clockwise or anticlockwise (as seen by a viewer towards whom the light is travelling), thus we might say (+)-limonene has the property of having the flavour of orange in combination with enough equivalent molecules, whereas (-)-limonene does not. Admittedly, we cannot in general read off the properties of molecules from those of substances they are made of (because the substances' properties are not necessarily homoeomerous). For example, mixtures containing equal quantities of each isomer don't rotate plane polarised light at all, but it doesn't follow that this is true of the quantities of each isomer when not mixed. In point of fact it isn't true of them, so pairs of optical isomers do always differ with respect to at least one property: the direction in which they rotate plane polarised light. This is very plausibly a property of the individual molecules themselves, not of substances made from them, so optical isomers do not share all their properties.

3.9.2 Premise Two

So far so good. What about the second premise, that optical isomers may have exactly the same parts? This cannot be well motivated for strict numerical identity between the parts, because the only way to show this conclusively would be to have a molecule which is in both configurations simultaneously, which clearly is not possible. So we must restrict the scope of our argument here to the philosophical principle that qualitative identity of parts entails qualitative identity of wholes.[40] With this in mind, we can justify the second premise.

As we have seen, the (+) and (-) molecules themselves are qualitatively non-identical as they have different physical properties: at a minimum, they rotate plane polarised light in different directions. However, the atoms which compose each optical isomer may themselves be qualitatively identical: they may have the same atomic number, atomic mass number, and so on. We would quite naturally expect qualitatively identical parts to have the same (in this case, actual physical) properties, and so – if we accept the qualitative extensionality principle – that the wholes they do compose should also be qualitatively identical. As we have seen in discussing premise one, they aren't.

In response, advocates of extensionality could claim that molecules have individual parts other than atoms – bonds, for example – and it is these parts which differ in cases of optical isomerism. After all, conversion from one configuration to the other requires the breaking of existing bonds and the forming of new ones, which just might be neither numerically nor qualitatively identical to those existing previously. On this view optical isomers share some parts but not others, which could certainly explain their differing properties.

But are bonds really parts of molecules? No doubt there is a sense of 'part' in which they are, but is this one which can be or should be accommodated by an extensional mereology? After all, bonds are one-sidedly dependent upon the atoms they constrain: there can be atoms without bonds, but no bonds without atoms. This follows from the composition of

[40] In this respect, our novel intended counterexample is actually weaker than some of the earlier ones, those which take two things to be made of the same parts simultaneously. However, as mentioned earlier the qualitative principle is perhaps the more philosophically interesting of the pair, so I think the weakness here is fairly slight.

atoms and bonds themselves. Atoms have as their parts electrons, neutron and protons (which have further parts themselves), whereas bonds are constituted by electrons only, and by those same electrons which are also parts of the atoms. In total, one oxygen atom and two hydrogen atoms possess ten electrons, while one molecule of dihydrogen monoxide possesses ten electrons *and* two bonds. If they are parts, what do the bonds add to the molecule? Several answers seem to be available here, dependent upon whether bonds are considered as discrete individuals or patterns of energy. Since this issue has yet to be settled by chemists, and is rarely discussed by philosophers (though see Hendry 2008), we should probably restrict ourselves to a few brief comments.

Perhaps extensional mereology could countenance bonds as parts of molecules which share all of their parts with, or 'overlap', the atoms, but given their respective compositions it is far from clear how this might accommodate the bonds being qualitatively non-identical when the atoms themselves are qualitatively identical, at least without some sort of double-counting of parts. And if the bonds are qualitatively identical in each isomer then the original problem remains. Lest it be thought that appeal can be made to *which* electrons figure in each bond, we should observe that it is not possible in general to individuate bonds according to which specific electrons constitute them, as in some cases the electrons may be 'delocalised' across a number of bonds and atoms.[41]

But whatever the promise of accommodating bonds in this way or any other, returning to more fundamental parts of molecules allows us to motivate more forcefully the second premise, in a way which avoids countenancing bonds altogether. Plausibly, the electrons, neutrons and protons which compose both optical isomers themselves, as well as their constituent atoms and bonds, may also be qualitatively identical in either configuration – they may have in each case the same mass, charge, spin and so on. In which case, optical isomers in either configuration have qualitatively identical parts, albeit not numerically identical ones.

[41] An example of electron delocalisation is to be found in benzene, which has a ring-like structure with bonds connecting its atoms of a length and strength between what would be expected for either single or double bonds. Benzene rings are commonly found in optical isomers.

3.9.3 Evaluating the Argument

So both premises of our argument for optical isomerism presenting a genuine counterexample appear to be true, at least when given some plausible assumptions and a weakened version of the second premise. Extensionality claims that identity of parts entails identity of wholes, so it seems that if this principle is true then we should expect that qualitatively identical parts of isomers (P2) always compose qualitatively identical whole molecules (P1). Since they don't, the argument runs, extensionality must be false in at least this case.

I imagine most readers will have smelt a rat by now, if not a couple of pages sooner. Whatever we make of the premises of this argument, it simply is not valid.[42] It only goes through if extensional mereology is supposed to tell us everything there is to know about the relationship between parts and wholes, but there is no compulsion to take the principle in either its qualitative or numerical form to be so ambitious. We can distinguish here between two distinct questions regarding the relationship between parts and wholes. First, van Inwagen's (1990) 'special composition' question of when some entities, the *x*'s, compose some other entities, the *y*'s; second the question of how the properties of the *x*'s are influenced by those of the *y*'s. Mereological extensionality imposes a restriction on what answers can be given to the first, in that they must be extensional, but it need not have anything to do with the second. So the extensional mereologist is free to respond to our argument as follows: the identity of a given whole object is determined by its parts quite independently of how they are structured or arranged, but the *properties* of the whole are influenced by the arrangement of its parts at any given time.

[42] I am indebted in particular to Jonathan Lowe, and also John Hawthorne, for helping me to see this.

3.10 Endurance versus Perdurance

All that is required to apply this response to the isomerism example is to believe that if a particular set of particles forms a (+) isomer at time t_1 and then a (-) isomer at t_2, the two molecules are numerically identical. That this is true follows from the unrestricted composition principle in classical mereology, which entails extensionality, and could also be accepted independently on a more restricted view (for example, we might take the view that particles only form molecules under certain conditions, but in those conditions the same particles always form the same molecule). This coheres rather well with the popular 'four dimensional' view, associated particularly with Quine (1960), (1981), Lewis (1986), and Sider, (2001), (2007b), that objects 'perdure' rather than 'endure' – they are composed of temporal parts, or 'stages', each of which is part of an object extended across space and time, often called a 'worm'.[43] But whatever the motivation, this is enough to render our isomerism argument invalid. Our (+) and (-) isomers can be numerically identical, yet have different properties because their (one and the same) parts are differently arranged at times t1 and t2. Phrased in perdurantist terms, the two isomers are temporal parts of one and the same whole (other perdurantists formulations are possible, for example by identifying material objects with stages rather than worms, as in Sider 2001, but this formulation is typical).

However, our argument for optical isomerism as a counterexample was restricted to qualitative identity of parts, and this response does not carry directly across to the qualitative extensionality principle; the parts of two isomers might be qualitatively identical yet numerically distinct. Just such a case would be where two (+) isomers transform into two (-) ones by swapping functional groups with each other (represented by *W*, *X*, *Y* and *Z* in figure 1), as on the perdurance view neither resulting molecule would be numerically or qualitatively identical to either of the earlier two. Nevertheless, essentially the same strategy can be consistently applied: composition itself is unstructured, but the way composite objects inherit their properties from their parts may be influenced by their arrangement. The space-time 'worms' which had (+) isomers as (all of their) temporal parts

[43] The cohesion is particularly intimate since stages are usually taken to be mereological sums.

at t_1 will at t_2 turn out to have rather different properties to (-) isomers, as our worms each span two molecules, but so be it. After all, if we are more interested in talking about the (-) isomers we can always switch our attention to the space-time 'worms' which include all the parts of each (-) isomer at t_2.

Of course, it doesn't follow from this that extensionality is uncontroversially true; there are other sound philosophical positions according to which it is false. For example, on at least some versions of the 'endurance' view according to which objects are wholly present at every time they exist, those objects are best characterised in terms of their identity and persistence conditions (see e.g. Baker, Lowe 1989, Wiggins). Thus, the same set of parts may compose two quite different things at the same or different times according to how those parts interact, and so composition is itself structured in at least some cases. Another option is to adopt a hybrid view according to which some objects endure but others perdure, which might or might not be compatible with extensionality as a universal principle, depending on the details. But the significant point is that our canvassed argument that optical isomerism constitutes a genuine counterexample fails.

3.11 The Futility of Counterexamples to Extensionality

In one respect this returns us to the position we were at the start of this section: bereft of any persuasive counterexample to either numerical or qualitative extensionality. But I think the preceding discussion enables us to make three significant methodological observations. Firstly, despite appearances, disputes over mereological extensionality do not turn on issues of ontological commitment to the sorts of entities which figure in various intended counterexamples. Optical isomerism as a counterexample has the same basic form as the existing counterexamples, that molecules exist is a relatively uncontentious claim (though see e.g. Rosen & Dorr, van Inwagen 1990), but as we have seen the argument still fails. The four dimensionalist strategy for rationalising extensionality with isomerism above does not involve denying the existence of anything, so the issue clearly is not one of ontological commitment.

The strategy works by characterising the objects which are uncontroversially (at least for present purposes) taken to exist such as molecules, cats and so on in a certain sort of way – for example as temporal parts of spatiotemporally extended objects – which is compatible with extensionality. Of course, taking on a very broad metaphysical doctrine just to preserve a more limited principle like extensionality would be putting the cart before the horse. But if you happen to think this is the best way to describe things like cats and molecules anyway, then you will be free to advocate extensionality if you want to. This is the second point, that our views on part-whole relations – and a great deal else besides – will be greatly influenced by how we characterise the objects the relations hold between.

In chapter two we discussed Poli's distinction between descriptive and formal ontologies, with the former providing raw material which is refined and systematised by the latter. The debate over mereological extensionality provides an excellent illustration of the methodological importance of this distinction. Once we realise that extensional mereology is based upon a perdurantist descriptive ontology, it becomes strikingly clear why repeated attempts to provide empirical counterexamples have failed: those material objects are differently characterised in the mereologists' descriptive ontology, in such a way as to prevent their being counterexamples. Any intended counterexample could only violate extensionality against the background of a particular conception of the objects which feature in it – a particular descriptive ontology – and we are free, subject to constraints of consistency and coherence, to conceive of them as we wish. It may be that mereological extensionality is false, in its numerical or qualitative versions, but this cannot be shown by pointing at things and exclaiming 'I refute thee thus!'.

Thirdly, making use of empirical data and examples is central to effective progress in formal ontology. This might seem an odd moral to draw since we have concluded that there can be no empirical counterexample to extensionality, but it follows from the discussion above. It was only through close analysis of actual cases of composition and parthood (with optical isomerism as the main example) that we were able to discern why the intended counterexamples failed, and that the disagreement over mereological extensionality runs far deeper than a mere clash of intuitions. This moral also contains a warning: careful study

of empirical data is an essential basis of the *a priori* reasoning so prevalent in ontology, but our formal ontologies cannot simply be 'read off' that data.

3.12 The 3D/4D Equivalence Thesis

A straightforward way to interpret this result is that endurantists may reject extensionality as a universal principle of composition whereas perdurantists will accept it as such. I think this is essentially right, and moreover very loosely reflects the current status quo. However in light of McCall and Lowe's (2003) 3D/4D equivalence thesis, which I am sympathetic to, this presents us with a slightly puzzling situation. According to that thesis, perdurantism and endurantism are equivalent in the sense that any claim made in the terms of one theory can be translated *salva veritate* into the terms of the other. Claims made according to one view are equivalent to claims made according to another in the same sense that formal systems can be equivalent despite choosing different primitives (McCall and Lowe make this point using point-based versus volume-based topologies, but we might equally use mereologies based upon improper versus proper parthood or versus overlap). What's more, according to the thesis there is no 'fact of the matter' in the world which makes one of the descriptions true and the other false (p.118), although for current purposes we don't need to agree with this additional claim.[44]

Assuming the truth-functional equivalence of endurance and perdurance views it might seem strange that different composition principles seem amenable in each case – structured composition for endurantism and unstructured for perdurantism. Which is the 'right' way to think about composition? Neither, or both? To properly assess the impact of the equivalence thesis on the structure of composition we ought to follow it through translating the sort of case we have been talking about. I am not sure I can provide anything as sophisticated as a full translation schema, but I will characterise a simple case of isomer

[44] Lowe's individual view differs slightly from that expressed by the 3D/4D equivalence thesis. If I have understood him correctly, he views the equivalence thesis to hold for perdurantists but not to hold for endurantists, hence providing an argument in favour of endurance (personal communication). For reasons of space we will limit our discussion to the version expressed by McCall & Lowe.

transformation in terms of endurance and the two principal perdurance views, and suggest how they may be intertranslated in this particular case.

As a simple example, take a single optical isomer and over a period of time rearrange its functional groups such that it is in the alternate configuration. A fairly typical endurance view of this would take the isomer to be wholly present at each point of its spatiotemporal career, but that career (and the isomer itself) only exists at times when all of the molecule's parts are suitably arranged. The two configurations' different properties arise from different arrangements of particles, and the molecule is clearly distinguished from the particles which compose it, whose own spatiotemporal careers are unbroken during the relevant period. At any time in the period there will be some fundamental particles, and at some but not all times they will also compose a molecule, in one isometric configuration or another.

On a stage perdurantist view such as Sider's (2001), what we call molecules are actually instantaneously existing objects, with one for every moment in our period of time. These stages are typically taken to be mereological sums (aka fusions), so there are a plethora of them at any time. To simplify things we will imagine there is only one – that made up of all the particles which are needed to compose our molecule. Series of different stages are related by temporal counterpart relations to form 'worms', which are derivative upon the stages they are made up of. Temporal counterpart relations are similarity relations, by analogy with Lewis' (1968, 1986) modal counterpart theory (which is discussed in connection with Casati & Varzi's mereotopology in chapter four). This allows the stage perdurantist to identify the molecule with the particles it is composed from, yet allow for future differences. Say our molecule is in a (+) configuration at t_1, is partially or wholly disassembled during the interval t_2, and in a (-) configuration at t_3. On this view we can take the molecule and particles to be one and the same object at every stage both exist, but different stages bear different properties according to – amongst other things – the arrangement of their parts. Thus stages during t_2 may be temporal counterparts to the stages at t_1 and t_3, but these stages are not molecules as the particles are not suitably arranged.

Worm perdurance differs from stage perdurance largely in the semantic claim that what we call molecules are identified not with stages but with worms, which therefore assume ontological priority over their temporal parts.[45] One and the same worm follows a spatiotemporal course which includes as temporal parts the (+) isomer at t_1, various states of arrangement of particles during t_2, and the (-) isomer at t_3. Again, differences in arrangement of the particles at different stages allow us to distinguish between when they form a molecule and when they don't, and which configuration it will be in.

3.13 3D/4D Translation

So how can these three views be intertranslated? Here is a reasonable suggestion: any claim about an enduring object can be indexed relative to a point in its spatiotemporal career, which is generally taken by endurantists to be an event. As such these careers have temporal parts, and may be characterised in terms of worms and stages by analogy to perdurance views of objects, with a worm equivalent to a full event and a stage equivalent to one moment of the event. Claims about the spatiotemporal careers of enduring objects, then, may be translated into those about perduring worms and stages, subject to the constraint that they be 'flattened' so that no two stages or worms may coincide in their entirety (full spatiotemporal coincidence entailing identity on the perdurance view). The reverse procedure translates from perdurance to endurance: conceptual or other distinctions between stages or groups of stages, in our example based on whether they rotate plane polarised light or not, are transformed into ontological distinctions between different sorts of objects (in our case groups of particles versus optically active molecules).[46] The resulting stage and worm analogues can be treated as moments and events in the objects' spatiotemporal careers, from which statements about the objects themselves can be made derivatively. The only difference in translating into stage versus worm perdurance

[45] A further difference is that in worm perdurance the stages are related by equivalence rather than similarity relations, but that makes no difference to us here.

[46] 'Conceptual or other distinctions' is admittedly imprecise, and would need to be cashed out explicitly for this to constitute a general translation schema. Doing so, however, would involve providing an answer to van Inwagen's (1991) general composition question 'when does something compose something else?', which is widely regarded as being nigh-on unapproachable. Hence my earlier doubt about providing a general translation schema.

lies in whether the claims predicated of the enduring objects end up translated into those predicated of stages or of worms.

I take this to be sufficient to show that translation *salva veritate* between the three views is possible for our simple example, and by extension for more complex ones involving multiple molecules exchanging functional groups. Extension to a good few other cases of composition should also be fairly straightforward. As such, this is at least a partial victory for the 3D/4D equivalence thesis, though being restricted to a particular example it falls far short of full vindication.

3.14 Structure in Composition

So what does this tell us regarding structure in composition? Recall that whether we take composition to be structured or not ultimately comes down to how we characterize the composite objects we are interested in, but there is good reason to think that the principal ways in which metaphysicians characterize objects are equivalent in the sense of being intertranslatable. Does it follow from this that structured and unstructured composition principles are also equivalent? I'm not sure that it does. If it did, this would seem to suggest that the very question of whether composition is structured or not is somehow ill-formed – that it involves some conceptual error – and I simply cannot see how that could be the case.

Better to view our results differently. It follows from the discussion above that claims made in terms of structured and unstructured composition should be intertranslatable, but it only follows given suitable adjustments in the descriptive characterisations of the composite objects. They need not, and indeed cannot, be intertranslatable directly and so it is questionable whether they are strictly equivalent. Nevertheless they can be intertranslated, so the choice between structured and unstructured principles of composition is not a substantive one, but is rather pragmatic based upon what descriptive characterisation of composite objects which is in place. We have seen endurantism to favour structure and perdurantism to favour its absence, but matters need not be so straightforward.

Unstructured composition is perfectly compatible with endurantism, although it may well be necessary to adopt one or more structured composition principles as well.[47] Perdurantism as presented above requires structured composition, but I know of no reason why alternative varieties could not exist which permit unstructured composition – whether such versions would have any advantages is open to doubt, though.

We are now in a position to answer the question we posed at the start of this section: is mereological extensionality true or false? The answer is that it is neither. It is pragmatically acceptable given suitable descriptive assumptions about the nature and identities of composite objects, and unacceptable under others. Since these very assumptions are themselves equivalent – they are different ways of saying the same thing – whether we accept or reject extensionality is in no way a substantive issue.

That said, the question remains of how we would be best served in characterizing composite objects and their compositional principle(s). Since it is closer to ordinary or common-sense discourse I shall adopt, where relevant, an endurance view of objects and correspondingly the view that at least some instances of part whole relations are structured. I would like to stress that this is purely for convenience; some suitably convoluted translation of our later claims in terms of perdurance and unstructured parthood is no doubt possible.

3.15 Unrestricted Composition

Unrestricted composition is often taken to be the most objectionable axiom of classical mereology, given that it entails the existence of both very large numbers of objects and objects of questionable ontological status. In this section we will see that these concerns are ill-founded. Neither the number of objects posited by unrestricted composition nor their unusual natures are problematic. We will also consider and reject two arguments made by Markosian (2007) in favour of restricting composition.

[47] Certainly, it seems that any endurantist will also need to accept a temporally indexed parthood relation (see e.g. McDaniel 2004 p.144 for a convincing argument to this effect; John Hawthorne also takes it to be true, based upon personal communication), although whether this qualifies as a structural relation I am not sure.

For all its considerable popularity in recent metaphysics, owing in particular to David Lewis, the idea that composition is unrestricted is extreme. If for two or more objects whatsoever there is a whole which comprises them as parts, then there are an awful lot of composite objects out there in the world, and many of them are deeply strange. My left toe and the moon form a whole object, on this view, as do Socrates and the Mary Rose. To borrow Varzi's (2010) terminology, any mereology including the general sum axiom – hence committed to unrestricted composition – seems to be both ontologically *exuberant* and ontologically *extravagant*. Exuberant because it posits an enormous number of objects, and extravagant because of their unusual natures.

3.15.1 Exuberance

There is a standard response to the charge of exuberance, which draws on comparing parthood to identity. Both identity and improper parthood are partial orderings – they are reflexive, antisymmetric and transitive – and as such the former can be considered to be a limit case of the latter.[48] This seems to invite Lewis' (1991) observation that parthood and identity are to some degree analogous. Furthermore, he argues, the analogy is close enough that the 'are' of mereological composition is a sort of plural form of the 'is' of identity. If this is right, then mereology is not so exuberant after all:

> Given a prior commitment to cats, say, a commitment to cat-fusions is not a *further* commitment. The fusion is nothing over and above the cats that compose it. It just *is* them. They just *are* it. Take them together or take them separately, the cats are the same portion of Reality either way.
>
> Lewis 1991, p.81

[48] This is not to suggest that the analogy between parthood and identity is a purely recent development. Harte (2002) claims, I think rightly, that Plato argues against composition as identity in his *Parmenides* and elsewhere.

> So striking is this analogy that it is appropriate to mark it by speaking of mereological relations – the many-one relation of composition, the one-one relation of part to whole and of overlap – as kinds of identity. Ordinary identity is the special limiting case of identity in the broadened sense.
>
> *Ibid.*, pp.84-5

But whether mereology is ontologically innocent in this way is doubtful to say the least; it is unclear exactly how parthood and identity are supposed to be analogous and in what way. As Lewis tells us elsewhere (1994) a raven is like an armoire to at least some extent, meaning that with all analogies the devil is in the details. In this case they aren't forthcoming; I cannot see why the analogy is 'striking', and Lewis does not make it clear why he thinks it is. Furthermore, as van Inwagen (1994) has pointed out, we already have a plural form of the 'is' of identity – the 'are' of identities between multiple things – so it isn't as if there is a vacancy here which needs filling. [49]

The analogy between parthood and identity is, I am inclined to think, both unclear and far from compelling. Admittedly improper parthood takes identity as a limit case, so it would be wrong to say the two are entirely disjoint – but so do all other partial orderings, and for that matter all equivalence relations such as connection or logical equivalence. I don't think anyone would seriously say that all such relations are analogous to identity in anything like the sense Lewis seems to have in mind, and I simply do not see on what grounds parthood stands out. Sider (2007a) talks about parthood being, like identity, an 'intimate' relation, contrasting both with causal and psychological relationships. This seems intuitively right, but falls far short of justifying a strong enough analogy between the two to support Lewis' claims quoted above. Particularly so since Sider's account of the intimacy of parthood is given in terms of a set of theses about its nature he finds intuitively attractive. I am much more inclined to think the difference in intimacy between identity and parthood on the one hand, and hitting and liking things on the other, lies in the former being formal ontological relations whereas the latter are not.[50] Admittedly, I cannot claim to be anywhere near as accomplished a philosopher as Sider, but if it comes to trading intuitions I doubt I am at any

[49] For a detailed and critical discussion of mereology's ontological innocence see Yi (1999).
[50] Incidentally – and this is something of a spoiler for a later argument – when I claim that representation is itself a formal relation there is no tension with this sentence. After all, I also claim that the representational theory of mind is false.

disadvantage to speak of. Philosophers' intuitions are notoriously corrupt; no doubt more so in proportion to their philosophising. For my part, I simply cannot find any clear argument, or follow any philosopher's intuition, which takes the analogy between parthood and identity to be any stronger than the description I have given here. There is one exception - Baxter's (1988) suggestion that parthood is *literally* identity. This should, I think, be quickly disregarded on the phenomenological grounds that it very clearly is not.

So is mereology ontologically exuberant? In the absence of a satisfying account of the analogy between parthood and identity, or any other strategy, I am inclined to think that it is, but also that this need not be seen as particularly problematic. As the evidence to support this view also addresses mereological extravagance, we will turn to that issue before discussing how to deal with both.

3.15.2 Extravagance

Many mereological sums will clearly be quite odd objects, spanning considerable quantities of space and time. However, we do not necessarily have to worry being committed to rather peculiar sorts of things existing. After all, mere existence doesn't entail that they actually *do* very much. If a mereological sum is something over and above its parts, and contra Lewis we have suggested above that it probably is, it does not follow that the mereological sum itself is terribly interesting outside of arcane philosophical contexts. In the absence of other restrictive commitments, why not countenance weird and wonderful objects like those made of bits of (whole) trout and bits of (whole) turkeys? Just don't imagine they are of any interest to fishermen or turkey farmers, or anyone apart from philosophers of a particular metaphysical bent (we might even follow Lewis 1991, p.213 in viewing these and other interests not in terms of ontological commitments but as implicit domain restrictions – much as in our discussion of the standard approach to mereological transitivity). Genuinely restrictive ontological commitments do, of course, come thick and fast for some philosophers, and an impressive range of restrictions upon composition have been suggested in the literature.

At the other extreme from unrestricted composition lies mereological nihilism, which takes there to be no composite entities at all. While classical mereology answers van Inwagen's special composition question (SCQ) 'when do the *x*'s compose something *y*?' by 'always', nihilism responds 'never'. Like Baxter's position, nihilism suffers from giving every appearance of being straightforwardly false, and at best it requires a rather elaborate semantics to explain all our talk of apparently composite things, which is not yet forthcoming (though for defence of nihilism see Rosen & Dorr 2003, Horgan 1993 and Unger 1979, 1980). Many more moderate positions have also been suggested in the literature, including some form of contact or fastening together (both strongly criticised in van Inwagen 1990); 'brutal' composition, according to which 'there is no true, non-trivial, and finitely long answer to SCQ' (Markosian 1998, 2007); and restrictions based upon particular sorts of things, most notably van Inwagen's (1990) criterion of being either a fundamental particle or engaging in activity which constitutes a life. A detailed survey and analysis of these options and others is given by Markosian (2007), to which I refer the interested reader. Each moderate position does face several challenges: restricting composition far enough, but not too far; avoiding other unintuitive or counterintuitive results; and satisfactorily answering both Lewis' (1991) argument that any restriction entails genuine vagueness in the world, and Sider's (2001) updated version that any restriction entails vague objects.[51] I do not know whether these charges can be met, but there is good reason to think the ontological extravagance of unrestricted composition unproblematic. The solution lies in returning to our notion of human interest.

There is a wealth of robust empirical evidence to show that from a very early age humans pay selective attention to certain types of objects and properties, to the exclusion of others. Michotte's landmark studies demonstrated that certain types of movement amongst objects in the visual field produce attributions of causality, such as 'the red ball is chasing the blue ball', while others do not. In every case, all that reaches the back of our eyes is colour and

[51] Koslicki (2003) gives what is to my mind a convincing rebuttal of Lewis and Sider's arguments. Lewis' formulation of ontological vagueness is seen to be based upon a vicious circularity, and Sider's argument rests not – as Sider claims – upon a dispute over whether key logical terms are vague or not, but on what the proper domain of existential quantification should be in the cases under dispute. As such, Sider fails to establish that there must be a determinate answer as to whether composition occurs, merely re-describing the same issue in non-mereological terminology (*ibid*. P.120).

movement at different times and places. We select salient aspects from these such as speed, timing, size and direction, and attribute causation on that basis. Somewhat similarly, infants from one month old react defensively to objects 'looming' in their visual field as if on a collision course by blinking (Shaffer & Kipp p.179), infants between 4-5 months can differentiate between objects on a collision course and those which will miss them (Schmuckler *et al.*), and infants between 3-5 months can differentiate between approaching objects and apertures and react accordingly (with objects and apertures represented by textural changes in the visual field; see Schmuckler & Li). Elizabeth Spelke and colleagues have demonstrated that young infants are sensitive to quantitative properties of objects such as number and relative size, but not to qualitative properties such as absolute size, shape or colour (we will return to Spelke's views in detail in chapter five, in relation to commonsense psychology). These results and many others tell us nothing about the number of objects which do exist, of course, but they do provide compelling reason to believe humans are naturally sensitive to certain types of properties and hence to those objects which exhibit them. This is highly conducive to a 'detectivist' view in ontology which hypothesises the existence of very many objects in the world, perhaps of very diverse types, and a human tendency to naturally recognise and respond to some over others. The ones we recognise and respond to are those which accord with our interests; from the evolutionary perspective appropriate to considering infant development the sole interest is survival. So we might well expect infants to be concerned to recognise visual cues which typically indicate an object on a collision course, and less interested in others. Given this empirical support for a detectivist view in ontology, we might well suppose that very peculiar objects exist – and perhaps in very large numbers. Just remember that they may only be of interest to philosophers.

That is not to say in any way that what actually exists is or should be relativised to human interests or concerns, simply that we have a built-in tendency to pay selective attention to certain types of properties (and so, by extension, entities which bear those properties) whilst selectively ignoring others. This does not entail an extravagant ontology, but is fully consistent with one.

3.15.3 Unwelcome Entailments

We have seen that there is strong empirical evidence to suggest mereological exuberance and extravagance are unproblematic. However, Markosian (2007) has recently pressed two further arguments against unrestricted composition, that it entails four dimensionalism regarding objects, and that it entails unrestricted diachronic identities between composite objects.[52] These two commitments would be substantial, and perhaps sufficient to reject unrestricted composition, but they can be disregarded since neither of Markosian's arguments is sound.

The first argument, adapted from van Inwagen (1990 pp.74-80), runs as follows. The particles which make up your body existed ten years before you did, and according to unrestricted composition they composed a whole object then. That object was not your body. Assuming extensionality, the arrangement of particles makes no difference to composition, so the older composite object still exists now.

> But it is impossible for two objects (such as yourself and the scattered object composed of the particles in question ten years ago) to become one. Which means that, according to [unrestricted composition], there are now two distinct objects located where you are located, each one of which is composed of the exact same particles…Now, the only plausible way to allow that two distinct objects can be in the same place, and composed of the same parts, at the same time is to say that the relevant objects, like two roads that share a stretch of pavement, are extended things that share a segment or "stage" or "temporal part".
>
> Markosian 2007, p.6 of online version

I am not convinced that *according to* unrestricted composition there are two distinct yet cohabiting entities, rather this possibility is not ruled out. Let's agree it is the case in this example. It simply does not follow from this that the two share temporal parts (hence four dimensionalism is true); there are other available accounts of cohabitation which are not implausible. In particular I think constitutionalism – according to which two objects *of a*

[52] A somewhat similar argument, based around the well-known paradoxes of material constitution, is given in section 3 of Sider (2007), but with the weaker conclusion that cohabitation (which Markosian takes to be entailed by unrestricted composition) favours four dimensionalism rather than entails it.

different sort such as a statue and a lump of clay may be co-located and composed of the same parts – is at least plausible, and moreover probably true. In a footnoted comment Markosian in fact admits that there are 'several ways of resisting a commitment to [four dimensionalism] that are available to the proponent of [unrestricted composition]' (2007 p.7, fn.17), which rather undermines his argument. As we are free to disagree that all of these ways are implausible, unrestricted composition does not literally *entail* four dimensionalism. This much should be clear from the fact that unrestricted composition is logically consistent with three dimensionalism (even ignoring the possible truth of McCall & Lowe's (2003) equivalence thesis); the only requirement is suitable assumptions about the identities of the two cohabiting objects.[53] For example, a three dimensionalist may consistently with unrestricted composition maintain that one of the objects in Markosian's example is a mereological sum (most likely the particles) and the other (your body) is not. Markosian's first argument, therefore, fails.

Markosian's second argument is adapted from one given by Sider (2001, ch.4 section 9), according to which proponents of unrestricted composition are committed to a diachronic version of the same principle:

> Unrestricted Composition with Unrestricted Diachronic Identity (UCUDI**):**
> Necessarily, for any non-overlapping xs, for any non-overlapping ys, and for any times, t1 and t2, such that the xs exist at t1 and the ys exist at t2, there is an object, z, such that z is composed of the xs at t1 and z is composed of the ys at t2.
>
> Markosian 2007, p.7 (formatting removed)

Pressed as a problem for unrestricted composition, this is essentially a strengthened version of the ontological extravagance objection: it turns out that we are committed to all sorts of deeply bizarre things, including now objects which exist intermittently (though see Simons 1987 for defence of just this view), and ones which change very radically over time. As such, we can give much the same response as to the original extravagance problem: so what? It's one thing to claim weird things exist, it's quite another to claim they are interesting or

[53] Much the same point has already been made by Varzi, who says 'there is no obvious reason why the endurantist should reject UC…There are, however, perfectly good reasons why the endurantist should better avoid UC' (unpublished 2006 p.6).

significant. I think this is ok. But another, stronger, response is available as well, one which picks up on an important difference between unrestricted composition and UCUDI.[54]

Unrestricted composition is a principle regarding when objects make up other objects, UCUDI is more than that. It adds to unrestricted composition the claim that any two composite objects existing at different times may be numerically identical. This is not a mereological principle, it is a principle about identity which is restricted to composite objects at different times. The closer we take the analogy between parthood and identity to be, the more pressing it will be to accept that unrestricted parthood entails unrestricted diachronic identity. But we have already said that the case for a close analogy is not compelling at all, even if some others find it so. And if we have learnt anything from our discussion of intended counterexamples to extensionality, it should be that issues of what we identify composite objects with should properly precede those of how we analyse composition, not follow them. Worse yet, perhaps, for Marksoian's argument is that it appears to commit a straightforward non sequitur: if x and y exist at distinct times then it follows that x+y exists, but it does not follow that x=y (this point was made to me by Peter Simons in conversation).

So we will conclude this section much as we began it: I see no reason to restrict composition, and know of no compelling argument to suggest that unrestricted composition is deeply problematic. The real potential pitfalls lie in what sorts of things we think mereological sums are, i.e. what we identify them with, and that is a different question altogether from whether they exist at all.

[54] Another response with which I have some sympathy, but do not endorse as such, would be to accept that identity is in at least some cases indeterminate, a possibility which has been suggested by Lowe (1998). This would block both Markosian's and Sider's arguments from unrestricted composition to UCUDI.

3.16 Mereology and the Three Problem Cases

It turns out that each of the five mereological principles we have considered can be maintained in the face of problematic examples and opposing arguments, at least under some small element of revision. Irreflexivity and antisymmetry are unproblematic provided we take care translating between proper and improper parthood, and transitivity can be rationalised with numerous apparent counterexamples by interpreting them as involving some restriction on mereological parthood. Extensionality turns out to be immune to counterexample, and by making suitable adjustments to the descriptive ontologies in use we can readily translate between extensional and non-extensional composition, making the choice a relatively minor and reversible one. Lastly, unrestricted composition was seen to be acceptable, at least in the absence of any independent and non-obligatory restrictive metaphysical commitments. It is not quite true to say with Lewis that classical mereology is 'perfectly understood, unproblematic, and certain' (1991, p.75), but we can at least conclude it is defensible in the face of opposition. So how well does classical mereology fare when applied to our three problem cases?

1. What is the difference between a cairn and some scattered stones?
2. In what sense, if any, has a car changed if we place its engine on the back seat?
3. In what sense, if any, is half of an uncut apple *part* of the apple?

Classical mereology answers questions one and three with 'none', as it takes composition to be entirely unstructured and so insensitive to arrangement of parts, and makes no distinction between detached and undetached parts in 2 for the same reason.
That is not to say adherents of classical mereology claim there is absolutely no difference in each of these cases – there patently is. Rather, they claim that there is no *mereological* difference, and hence no difference solely in terms of parts and wholes.
On this view, part-whole relations are oddly circumscribed; if that is how some philosophers want to think about them, so be it. But parts and wholes are in general parlance much richer concepts, and ones which clearly require a more sophisticated set of resources to do justice to than classical mereology alone, as these and innumerable similar examples make very clear.

These demonstrate that mereology does not have the resources needed to capture the distinctions made in each of our problem cases. The response that these differences are non-mereological is fine if by 'mereological' we mean 'addressed by classical mereology'. That much is obvious. But if by non-mereological we mean not a question of composition, this is clearly false. Classical mereology alone is inadequate as an account of any of our three problem cases.

To abandon mereology, however, would perhaps be premature. As we have seen its main principles are fairly robust, and it is closely connected with at least some formulations of four dimensionalism. Whether this is a good thing or not depends upon who is asking, I suppose, but given the popularity of four dimensionalism it would be unwise to drop mereology too quickly. Another less revisionary response would be to ask how mereology can be supplemented to deal more effectively with these cases, and others like them. Just this strategy is taken up by Casati & Varzi (1999) and Smith (1996a, 1996b), who supplement classical mereology with a topological primitive and axioms to model a range of geometric concepts including contact, connection, boundary and interior. The strategy is not new – Simons (1987) notes that 'the introduction of topological notions is in some sense the natural next step after mereology' (pp.92-3), and that nearly half a century earlier Menger (1940) suggested topology be given a mereological rather than point-set theoretic basis. We will concentrate upon Casati & Varzi's discussion, as it is the most systematic and developed philosophical treatment to date. In the following chapter we will evaluate how successful their strategy is in tackling our three problem cases, before comparing the attempt with Kit Fine's rival theory of parthood.

4. Topology and Non-mereological Composition

4.1 Topology for Philosophers

In tackling our first example, the shortcoming of classical mereology is that it makes no distinction between wholes which are scattered and those which are not. Plausibly, a cairn only exists when the stones it is composed from are arranged sufficiently closely (having accepted unrestricted composition we must maintain that a mereological sum of these parts exists come what may, but it does not follow that that sum must be a *cairn* come what may). If topology can be used to model a suitable concept of connection, then intuitively it should prove very helpful in tackling this example – to a first approximation, we would be able to say that a cairn is a group of connected stones, *modulo* some additional constraints. It may also be of use in the other three examples. In this section we will explore to what extent topology can usefully supplement mereology, and apply the resulting 'mereotopological' system to our four examples. Recall that we saw in discussing optical isomerism on page 89 that, since different modes of spatial orientation exist in all non-zero dimensions, spatial categories *must* be used to complement mereology in any formal ontology of objects in one or more dimensions. Our question here is whether topology can provide the appropriate spatial apparatus to analyse our problem cases, each of are three dimensional.

Topology is 'the mathematical study of the properties that are preserved through deformations, twistings, and stretchings of objects' (Weisstein), and provides a qualitative complement to geometry's quantitative study of space and spatial properties. As such, topologically equivalent shapes can vary quite substantially in other ways. A sphere is equivalent (homeomorphic) to a cube, a cylinder or any other shape it can be continuously deformed into. Poking a hole in it, however, produces a topologically distinct shape – a doughnut, or torus – which is homeomorphic to various other shapes, famously including coffee mugs. A second hole separate from the first produces a new topology again, and an extra degree of complexity arises when the resulting loops are threaded through each other,

as in a pretzel. Three particularly interesting topological shapes are the torus, Möbius strip and Klein bottle.

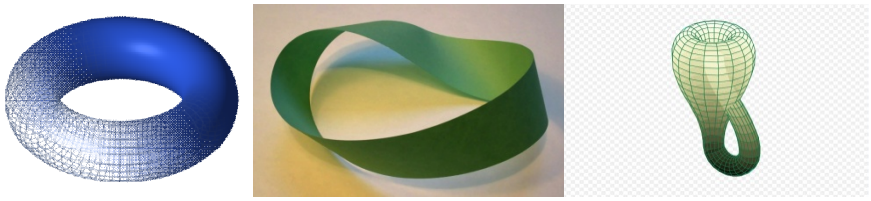

Figure 2:[55] Torus. Möbius strip. Klein bottle.

Each shape has a single surface, with the Möbius strip having one edge and the Klein bottle no edges at all (and when produced in four dimensions so as to have no hole in its side, it has no interior or exterior either). They each have a number of interesting properties; for example the four colour theorem does not apply to a map projected onto any of them. The theorem states that any map may be coloured in such a way as no adjacent sections are the same colour by using just four colours, but only applies to maps on a plane (a flat 2D surface). A map projected onto a torus requires no more than seven colours, and for a Möbius strip or Klein bottle up to six are needed (if this seems strange, imagine forming a torus from a sheet of paper by rolling it into a cylinder then joining the cylinder's two ends, or twisting a map and attaching the two straight ends to form a Möbius strip. Klein bottles are harder since they can only be properly made in four dimensions). Another useful application of topology comes from the three utilities problem: three houses need to be connected to three utilities companies. Can this be done without any of the supply lines overlapping?

[55] These images are reproduced from *Wikipedia Commons* (URL=http://commons.wikipedia.org) under the terms of the GNU Free Documentation License 1.2.

Kuratowski's theorem (1930) proves that the answer is 'no', at least on a plane. In the resulting figure there will inevitably be some lines which cross each other, looking something like this:

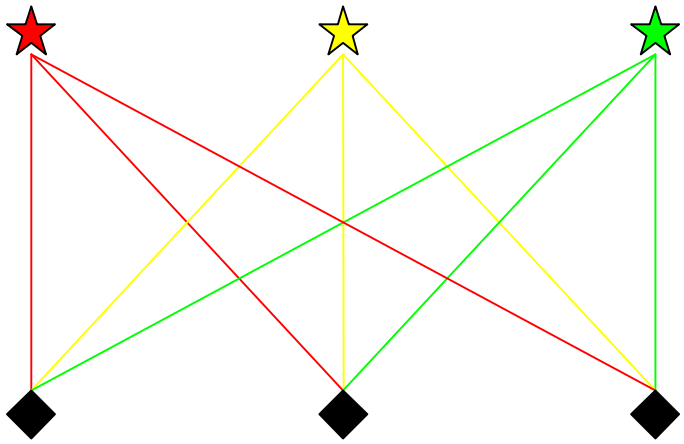

Figure 3: Three utilities and three houses on a plane.

When projected onto a torus, however, the problem can be readily solved (dotted lines are on the far side of the surface):

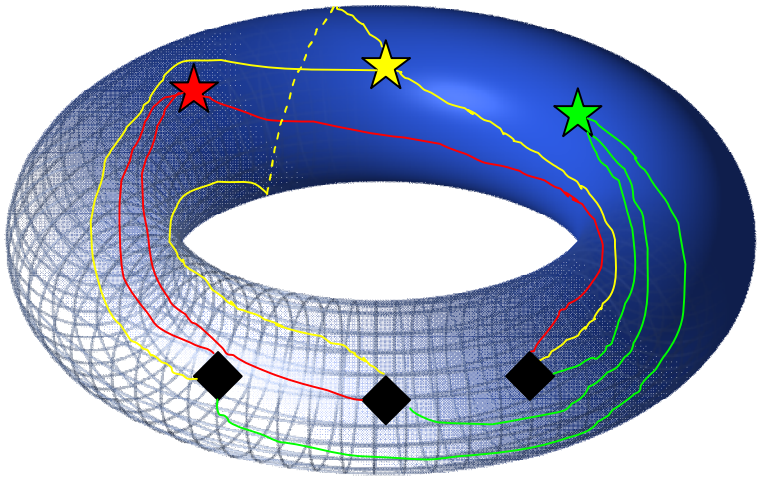

Figure 4: Three utilities and three houses on a torus.

Topological shapes have more practical applications as well (although this result could perhaps be some use to circuit board designers); the uteractive ingredient in Kalata-Kalata, a traditional African brew given to women during labour, is a protein from the plant *Oldenlandia affinis*, which is homeomorphic to a Möbius strip. I have heard that many conveyor belts and old-fashioned printer ribbons are made as Möbius strips to ensure even wear – even if this is somewhat apocryphal, the principle seems sound enough. Moreover, topology is fundamentally a study of spatial concepts so it should be well suited to being applied to things which occupy space, a point not lost on Casati & Varzi (1999) – nor should it be here since each of our three problem cases involves essentially spatial objects. So how does mereotopology – mereology supplemented with topology – actually work?

4.2 Mereotopology

Casati & Varzi (1999) distinguish three ways of relating mereology and topology: subsuming one under the other, doing so vice versa, or taking one primitive from either. They favour the latter choice. Mereotopology could be easily viewed as a restricted form of mereology – unrestricted composition tells us that any parts compose a whole, whereas adding topological elements allows the formulation of less exuberant principles, for example that only parts which are connected to each other form wholes, or only parts whose interiors are connected form wholes. On this reading, adding topology is equivalent to a φ-restriction, as made use of by the 'standard account' in chapter three's discussion of the transitivity of parthood. This approach would only require the use of a single mereological primitive. Conversely a system could be worked out based upon a single topological primitive, with mereological parthood acting as φ-restriction upon it. Casati & Varzi (1999) explicitly rule out both of these options, and instead take mereotopology to require two distinct primitives. As I see it, what they have in mind is that mereology and topology are complementary, one providing the resources to talk about parts and wholes and the other to talk about connection and other spatial concepts. Together, they provide a richer framework for discussing part-whole relations than mereology alone, and remain largely if not entirely topic-neutral (see below). It is this approach which we will examine.

As with our presentation of mereology, we will restrict ourselves to the main principles of the system, and skip over other theorems and corollaries which are interesting but peripheral to our purposes. Casati & Varzi (1999, ch.4) give the most thorough and detailed presentation I know of, and again I see no reason to reproduce their discussion any further than necessary. A little more so than mereology, the formal mereotopological apparatus involves a moderate degree of technical sophistication, but again this can be separated from the conceptual significance of the principles involved. Naturally enough, mathematical topology includes numerous distinct branches, all with a great deal more technical sophistication than the present discussion of mereotopology, or any in the philosophical literature. But so be it; it turns out that where mereotopology fails us, it is its basic spatial concepts which are inadequate for our purposes, not the details of how they are expressed.

Taking C*xy* to represent '*x* is connected to *y*', assuming predicate logic with identity and classical mereology, and suppressing initial universal quantifiers, Casati & Varzi's mereotopology can be developed as follows:

C1	Reflexivity	C*xx*
C2	Symmetry	C*xy* $\rightarrow$ C*yx*

As a similarity relation, connection will have to be reflexive and symmetric but is non-transitive. It is related to parthood by C3:

C3	Everything which is connected to a part is connected to the whole.	P*xy* $\rightarrow \forall z$(C*xz* $\rightarrow$ C*zy*)

It is helpful here to add two mereological definitions which were omitted for simplicity in chapter three, those for summation and general summation. Taking 'ι' as a description operator (Russellian or otherwise), we have:

$x + y =_{df} \iota z \forall w(Owz \leftrightarrow (Owx \vee Owy))$ Summation

$\sigma x Fx =_{df} \iota z \forall y(Oyz \leftrightarrow \exists x(Fx \,\&\, Oyx))$ General Summation

With these in mind we can define the property of being self-connected, which obtains for a whole when any two of its parts are connected, and also the concepts of internal part and interior, complement and exterior, closure and boundary, which will be needed in our discussion:

$SCx =_{df} \forall y \forall z(x = y + z \rightarrow Cyz)$	Self-Connection
$IPxy =_{df} Pxy \ \& \ \forall z(Czx \rightarrow Ozy)$	Internal Part
$ix =_{df} \sigma z IPzx$	Interior
$\sim x =_{df} \iota z \forall w(Pwz \leftrightarrow (Pwx \ \& \ \neg Owy))$	Complement
$ex =_{df} i(\sim x)$	Exterior
$cx =_{df} \sim(ex)$	Closure
$bx =_{df} \sim(ix + ex)$	Boundary

Intuitively, self-connection can be used to distinguish between scattered and non-scattered wholes as it should not be satisfied by the former. However this is not quite right as any series of objects in barest contact would satisfy SC. Classical mereology tells us that they do form a whole, but we have seen that the concepts of part and whole in mereology *simpliciter* cannot do justice to three problem cases. What we need is some way to distinguish between all of the wholes countenanced by classical mereology and just the more 'natural' unified ones such as tables, trees, bodies and so on. Self-connection goes some way towards this, but is clearly not strong enough. With this issue in mind, Casati & Varzi propose two stronger predicates, strong self-connection and maximally strong self-connection:

$\mathrm{SSC}x =_{df} \mathrm{SC}x \,\&\, \mathrm{SCi}x$	Strong Self-Connection
$\mathrm{MSSC}x =_{df} \mathrm{SSC}x \,\&\, \forall y(\mathrm{SSC}y \,\&\, \mathrm{O}yx \rightarrow \mathrm{P}yx)$	Maximally Strong Self-Connection

Strong self-connection rules out wholes whose interiors are not all connected, but is still satisfied by internal parts of larger wholes, such as the core of an apple. Maximally strong self-connection rules out these sorts of cases, by requiring that the whole it applies to includes every strongly self-connected whole it overlaps. In other words, maximal wholes are the largest strongly self-connected ones.

An example might help illustrate these two predicates. Imagine a sphere of homogeneous matter ('gunk') of diameter *x*. The sphere of matter of diameter *x-n* for some positive *n*, which shares a central point with the original sphere, would satisfy SSC but would not satisfy MSSC. It would not satisfy MSSC because it overlaps other portions of matter smaller than the entire sphere without them being entirely contained within our internal sphere. In other words, there is a *remainder* of matter left over if we subtract our SSC-satisfying sphere from the object we intuitively want to call the whole. This object, the intuitively whole original sphere, satisfied MSSC.

With one further refinement, this completes Casati & Varzi's mereotopological system (Smith's 1996a system is closely similar, but uses IP as its topological primitive)[56]:

$\varphi\text{-}\mathrm{MSSC}x =_{df}$	Relative Maximally Strong
$\varphi x \,\&\, \mathrm{SSC}x \,\&\, \forall y(\mathrm{SSC}y \,\&\, \mathrm{O}yx \rightarrow \mathrm{P}yx)$	Self-Connection

MSSC needs to be parameterised for at least two reasons. The first is technical – on the theory described above everything is connected to its complement so the only available candidate for MSSC is the universe itself; more broadly it follows from this that connection *simpliciter* is unhelpful for distinguishing objects from their surroundings. The second is

[56] The complete system does also include analogues of the three Kuratowski (1922) axioms for topological closure (with P comparable to set-theoretic inclusion and + comparable to union), but these are omitted as they have no real impact on our discussion. See Casati & Varzi (1999) p.59 for details.

related to the first and is tied to mereotopology being, like mereology, both a formal and a formalised theory. It seems quite strange to take a system based on the specifically spatial concept of connection to be formal in the sense of cross-categorial or perfectly general, because there are at least some categories where spatial concepts simply do not and cannot apply. Universals, for example, plausibly do not have any kind of topology and – in contrast with parthood – it sounds suspiciously like a category error to ask whether they are connected. Since Casati & Varzi explicitly reject the idea that mereotopology is not a purely formal theory in the sense which mereology is – by insisting that their primitives for parthood and topology be both unrestricted – they need to limit topology in some other way, such as by parameterising their account of mereotopological wholeness. I suppose this is not too major an issue for us here, though, since each of our three problem cases concern objects which are necessarily located in space. As we shall see, mereotopology fares better than mereology alone as it can account for one of the three cases. To explain the other two yet more conceptual resources will need to be added.

4.2.1 Cairns and Connections

1. What is the difference between a cairn and some scattered stones?

How well does the mereotopological concept of maximally strong self-connection apply to the first problem case, of characterising the difference between a cairn and some scattered stones? At a first glance not well at all, although with some further refinements things look much more promising. For the sake of convenience, let's factor out some plausible constraints on what makes something a cairn – that it be located on top of a hill or mountain, and that it be man-made one stone at a time – and just think of them as piles of stones. Piles of stones do not in general satisfy MSSC or even SSC (although some piles might satisfy SC, depending upon their arrangement), because the interiors of the stones which make them up are not connected to each other. The parts of the pile which are stones are connected, but the interiors of those parts are not. Even piles of stones which tessellate perfectly, such as bricks, will not satisfy either principle because the sum of all the internal parts of the stones is not self-connected. The bricks are in contact, not continuous,

with each other. To satisfy MSSC the stones would need to be fused, for want of a better term, in such a way that all of the tangential (that is, not internal) parts of the individual stones which were not also tangential parts of the pile itself should become internal parts of the pile. To clarify this, compare the diagrams in figure 4 below (the gaps between the stones are not necessary for the point being made, but make illustrating it easier – without them *B* and *C* are the same):

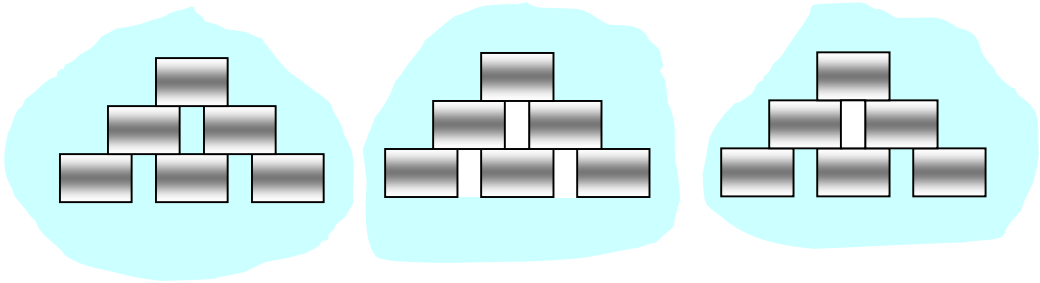

Figure 5: *A* *B* *C*

In each diagram the local exterior of the pile is represented in blue, the stones are grey, and in *B* and *C* the white areas within the diagrams represent parts of the interior of the pile which are not occupied by a stone or part thereof. *A* represents the model provided by Casati & Varzi's mereotopology.[57] *B* and *C* are alternative models, where the boundary of the whole pile does not correspond exactly to the sum of the boundaries of the stones – in this sense that the boundary may exceed the sum of its parts. We will discuss the distinction between boundaries which do, and those which need not, follow physical continuities (*bona fide* and *fiat* boundaries respectively) below, but accepting it as unproblematic for the time being we seem to have a promising line of approach to our pile of stones if we accept that the pile has a *fiat* boundary demarcating its interior. As we shall see, Smith and Varzi (2000) argue that *fiat* boundaries such as those in *B* and *C* can be modelled by mereotopology, providing a formal theory according to which they are (in contrast to ordinary point-set

[57] Bearing in mind the distortions permitted by topology, there will of course be numerous other models which are quantitatively quite different, but topologically and mereologically equivalent. Such distortions would have no effect on the argument here.

theoretic boundaries) capable of overlapping each other. This seems right to me, and I shall simply take it to be so, deferring further support of the claim to Smith & Varzi's article.[58]

A further issue still remains. Topology allows significant 'deformations, twistings, and stretchings', but piles of stone do not: some configurations are stable whereas others are liable to collapse, for example one obtained by turning one of the pyramid-like piles in figure 4 on its head. We might end up with a pile of stones (and very plausibly the same pile at that), but equally the parts of the pile could scatter widely enough that there is no pile, and so no cairn, left at all.

This suggests that in order to do full justice to the intuition that a pile of stones (or a cairn) is quite different to some scattered stones, we need to supplement mereotopology with some additional factor, to capture the dependence of the whole upon not just (1) the existence of its parts and (2) their suitable topological connection, but also upon (3) the causal interactions between them which maintain a stable configuration, at least for a time. A stable configuration of a composite object would, in general, be one where the object's causal powers do not vary to a significant degree.[59] For a pile of stones, causal stability involves not moving much at all (relative to the surface it rests upon), but this does not mean being inactive: 'just sitting there' still involves causal interactions, as unmoving objects are still acted upon by various causally potent forces such as gravity.[60] Instead of causal stability, then, we might equally talk of causal equilibrium; for present purposes I shall take the two to be the same. It seems very plausible to me that, as long as the boundary around the pile of stones is a *fiat* one, the first problem case can be tackled effectively by using mereotopology with an extra factor along the lines of 'causally stable' as a restriction upon MSSC. No doubt the restriction is only suitable for some spatial entities, and could be cashed out more precisely by someone with a better knowledge of the

[58] See also Smith & Varzi, 2001 for further discussion and application of the distinction to organisms and their environments

[59] I do not say that its causal powers do not vary at all, because a system which varies within limits might also be described as causally stable. For example, in regular humans blood glucose levels are regulated by the Islets of Langerhaans in the pancreas, which produce glucagon and insulin to raise or lower blood glucose levels respectively. It certainly seems reasonable to describe this system as causally stable, at least when functioning normally.

[60] I'm not sure whether in the absence of any gravity, or any other attractive or repulsive forces, the same stones in the same spatial configuration still count as a pile, but I'm inclined to think not.

physical sciences (and we should not forget our factored out constraints as well), but I take it that this schematic approach to accounting for the unity of a cairn is sufficiently promising that we can leave the rest of the details for another time. These will no doubt include a requirement that the cairn be deliberately made by human hands, and the geometrical constraint of being broadly conical with the apex at the top. Neither of these can be expressed using topology, but there is no reason to think they cannot be added to the mereotopological apparatus used here.

There are also alternative topological concepts which could be brought to bear on this example. Simons favours path connection over MSSC, on the grounds that the former is a more intuitive concept (this point was made in conversation). The idea is that two objects are path-connected if we can imagine a line which passes through both without any part of the line being outside one or the other object. This is straightforward to apply to a pile of stones, and models C in figure 5 without need for any further conceptual apparatus. Furthermore all path-connected spaces are connected but not vice-versa, so it is a more precise account than Casati & Varzi's. For these reasons Simons is correct to say it is preferable for our example. For continuity of presentation with Casati and Varzi, however, we will continue to use MSSC as a benchmark for topological connection. Since all cases of MSSC are *ipso facto* path-MSSC all our results can easily be carried across to the more precise concept. We will return to *fiat* and *bona fide* boundaries below, but here our discussion of the first problem case draws to an end.

4.2.2 Cars, Connections and Misplaced Engines

2. In what sense, if any, has a car changed if we place its engine on the back seat?

What of our second case, where a car's engine is removed and placed on the back seat? First of all, a tacit assumption: after being moved the engine is no longer connected to the fuel intake or exhaust in such a way that it can draw fuel and emit waste gases from and to them, and cannot transmit power to the wheels. Otherwise, we might well be inclined to

say that nothing of any real significance has changed here, passenger comfort notwithstanding. But given this assumption, mereologically speaking still nothing has changed in this situation, nor has anything changed topologically (at least when comparing the original and end state – if the engine was fully separated from the car at any point then we would have a violation of MSSC for the car. We are only really interested in the original and end states though, and can easily assume that the engine remained suitably connected throughout). But surely there is an important difference in our problem case: the car still retains all of its earlier parts, and in a topologically equivalent arrangement, but it no longer works. It seems clear that, as with our first problem case, mereotopology alone is insufficient to capture the distinction we intuitively want to make here. We shall see that in the second case we can also construct a plausible schematic account around a mereotopological core.

Cars are in some respects a lot like piles of stones: they are physical objects, artefacts, and are made from various parts which are capable of existing independently of the whole they compose. Just as a pile can be scattered without destroying the stones, a car can be disassembled without destroying its component parts. Likewise, both piles and cars require a certain degree of causal stability, or equilibrium, between their parts and environment to maintain their existence. As such, I take it that the account sketched above for piles of stones will also apply to cars; any car needs to have its parts connected together, they must all be within a *fiat* boundary demarcating the car, and be in a configuration which is causally stable. A car needs more than this, however: it also needs to work. Temporarily inactive but workable cars are still cars though, and plausibly broken cars are as well (provided they are not too badly damaged) so we might want to relax this criterion a little to allow a car to be something which *can* work as a car, or has worked while in a relevantly similar state.

Likewise an engine: conventional combustion engines are non-scattered composite objects, ones which produce motion power by burning fuel. In other words, they work.[61] Vieu & Aurnague (2007) impose a stronger restriction upon what it is to work, or function, as an engine:

> In addition, we want to characterize the full function, and not the local behaviour. For instance, we do not consider as being functioning an engine switched on and "working" flawlessly, i.e., producing motion power, while detached from any machine; we want to qualify as "being functioning as an engine", an engine working flawlessly within the larger system of some machine using its motion power.
>
> p.321

On this 'full function' view the salient difference between the original and later configurations of the car and engine is that as well as the engine changing location, it no longer *works as part of* the car. This seems intuitively right: the rearrangement which has taken place has neither destroyed the car nor the engine, and the engine remains part of the car afterwards as well as before. However, when sitting on the back seat the engine is not a *working* part of the car.

Some might think it a little too quick, however, to say that the rearrangement has not destroyed the car. According to mereological essentialism, defended most prominently by Chisholm (1975), nothing may survive a change of parts. More precisely,

> for all objects x, for all objects y such that y is part of x at some time, in every world in which x exists, y is a part of x at all times in that world that x exists.
>
> Cameron 2009, p.1 of online version

According to this doctrine, the act of removing the engine did destroy the car (although there would still be something else made of the rest of the car's parts, which we might *call* the car). Given this, when the engine is placed on the back seat, we are presented with a range of options:

[61] I don't mean to suggest that in order to work a part must itself be composite – one counterexample to this claim is the crude compass made by floating a bar magnet on water – but mechanical examples will largely tend towards complexity.

a) We could deny that this constitutes any genuine car at all since the resulting 'car' doesn't work (although on our more relaxed criterion for work this option may be ruled out as the 'car' is close to a working state).

b) We might admit that reintroducing the engine does produce a car, and one with all the same parts as the original one, but nevertheless the two are not the same (perhaps because one works while the other does not, or because of a general suspicion of intermittent existence).

c) We might be inclined to accept intermittent existence and take it that the same car exists at the start and end of the engine's being moved, but not necessarily in the middle (for defence of this view see Simons 1987 ch.5).

Other responses are available as well. We can remain neutral here regarding each of these positions, and mereological essentialism itself, by simply assuming that the engine remains suitably connected to the car throughout its movement. A suitable connection would be one which satisfies the schematic account of wholeness sketched out for the first problem case, according to which a car is effectively a pile of car components, albeit one which cannot vary too far from the configuration in which it can be driven. This is all rather vague, I know, but then again persistence conditions for artefacts very probably are vague themselves. There is plenty of fruitful discussion to be had over mereological essentialism, but it need not worry us any further here.

To return to our main point, the second problem case clearly requires some notion of working, or equivalently of functioning, in order to differentiate between a car with its engine in place, and when the engine has been removed and placed on the back seat. A car must be more than just a pile of car parts, otherwise there would be no difference between the two. The difference is that the car must be able to work, or be close enough to working to count. A full explication of the notion of working, or functioning, is far beyond the scope of this thesis, nor is it necessary for our purposes. Our aim is to first assess the extent to which mereotopology can tackle the problem case, and second, given that it is lacking, to give a schematic account of how it might be improved. Mereotopology cannot differentiate between a car and any other connected arrangement of car parts; what is required is an

account of what it is not just to be a connected part, but a working part as well. We will conclude this section with a brief survey of some ways to cash out this notion.

In recent decades the majority view of function in philosophy seems to be aetiological, taking functions to be teleological (goal-oriented) on the one hand and reducible to natural selection on the other.[62] Perhaps the most developed version of this view is Millikan's theory of proper functions (1984, 1989, 1995, 2005) which we discussed in chapter one. To recap, a mechanism or process' proper function is the way it works under normal circumstances, and so proper function may diverge quite substantially in some cases from how something actually works. In contrast to this neo-teleological view, Cummins advocates what he calls functional analysis which does not address a 'why-is-it-there question', rather a 'how-does-it-work question' (2002, p.2 of online version). In particular, 'to ascribe a function to something is to ascribe a capacity to it which is singled out by its role in an analysis of some capacity of a containing system' (1975, p. 765). While Cummins' analysis is more immediately germane to the discussion above than Millikan's I won't try to compare or adjudicate between them (for comparisons and criticisms see Cummins 2002 and Millikan 2000b, 2002a). There are of course many other theories available, not all of which fall on either side of this divide. Dennett (1995) proposes a teleological theory close in many respects to Millikan's, while Johansson (2006) expresses sympathy with Cummins' approach, but suggests that function is ultimately a 'primitive undefinable concept' (p.4 of online version).

I don't much mind which of these theories is the strongest, either in general or applied only to our second problem case, although I am most sympathetic to Johansson's. If function really is a primitive indefinable concept then a straightforward way to analyse working parthood is to supplement mereotopology with a third logical primitive, say F*xy* for '*x* contributes to the functioning of *y*'. I leave the details of such a mereotopoloergonology for another time, and a better neologism as well. Although rumour has it that successfully being awarded a PhD requires making at least one original contribution to knowledge in the field – perhaps this could be mine? Anyway, taking this approach would require that either the

[62] For a good detailed overview of philosophical theories of function see Cummins *et al*. 2002.

theory of functions could be expressed in predicate logic with identity, or else that mereotopology be translated into some other logical system. In principle this seems reasonable, and the former strategy has already been put into practice. Garbacz (2007) presents a theory of functional parthood in predicate logic, but rather than adopting a separate primitive for function he defines a single primitive for 'functional part' which is, like mereological parthood, a partial ordering. As such, his approach is better interpreted as offering a restricted version of mereology (and so, by extension, of mereotopology). I am slightly suspicious about its suitability in general, given that Garbacz sets up his theory of function in terms of artefact designs, but I have no argument against this strategy to hand at the moment.[63]

We can see that mereotopology alone is inadequate to tackle either of our first two problem cases. The prospects for producing plausible responses by restricting or supplementing mereotopology are reasonably bright. In the first case, mereotopology is insufficient to characterise a cairn as a pile of stones since piles do not in general satisfy MSSC or SSC (and rarely SC); however, by first taking the boundary demarcating the interior of the pile to be a *fiat* boundary – one which does not necessarily follow the boundaries of the stones from which the pile is composed – and by stipulating that the pile must be causally stable, we found that in general a pile of stones may satisfy a restricted version of MSSC, whereas scattered stones do not. In the second case, we saw that accounting for the difference between a car with its engine in the usual place and on its back seat also requires a notion of functional parthood. This could be either separate from, or could incorporate the causal stability requirement in the first case, and could either be provided by introducing a third logical primitive, or else restricting mereological composition or MSSC to working parts and wholes. We have not provided a full theory of parts and wholes for either case, just noted some plausible conceptual requirements for such a theory.

[63] Incidentally, to refer back to our discussion of transitivity, if functional parthood is treated here as a monadic predicate then Varzi's interpretation of the transitivity of at least some problem cases involving functional parts cannot work. My interpretations of the examples a) and b), however, would stand either way.

4.2.3 Actual and Potential Parts

3. In what sense, if any, is half of an uncut apple *part* of the apple?

First off, let's assume that half of an uncut apple actually *is* part of the apple in some sense yet to be made clear. Second, since it is not specified how the half is to be measured let's assume there is no physical discontinuity marking the division – this distinguishes between an uncut apple and two apple halves placed next to each other. Third, we will take it that the half is strongly self-connected, to rule out arbitrary scattered sums of apple parts which add up to half its mass and/or volume (thus the half of the apple is the left or right half relative to some plane which passes through the apple's center, or else some other quantity of apple such as those in figure 5). Our uncut half will not satisfy MSSC because it is part of a strongly self-connected whole – the uncut apple itself – which does satisfy MSSC.

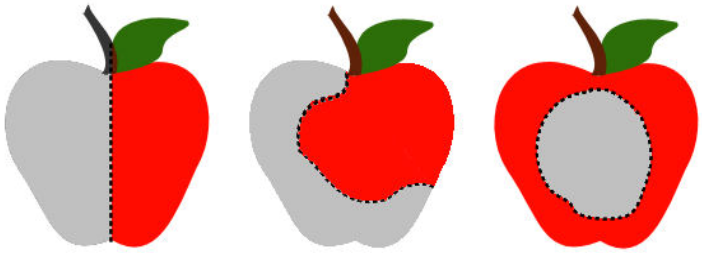

Figure 6: Three strongly self-connected (approximate) halves of uncut apples, with interior boundaries marked by dotted lines.

This suggests a straightforward analysis of our third problem case: a whole uncut apple satisfies MSSC, as do both halves following its being cut. Prior to the cut, both halves only satisfy SSC. This gives a precise answer to our question – a half of the uncut apple is a part of the uncut apple which satisfies SSC but not MSSC, and has a volume or mass (or both) which is 50% of the apple's.

This answer reflects a traditionally – and also currently – popular viewpoint which, following Holden (2004), we will call the actual parts doctrine (AP). According to this view, the parts into which any composite object may be decomposed each exist independently of both the whole and of the other parts, and it is clearly implicit in classical mereology (the actual parts

doctrine is a more precise formulation of what we called 'bottom-up' and 'atomism' in discussing mereology's descriptive ontology in chapter two).[64] Actually, to call it popular may be something of an understatement – in an overview supported by numerous quotations Holden (pp.86-7) attributes this view to Descartes, Newton, Galileo, Bayle, Berkeley, Wolff, Hume and the early Kant, amongst others. I suppose its popularity was enhanced by the widespread appeal of atomistic science from Galileo and Descartes onwards, but it is important to remember that AP is a metaphysical theory which is quite independent of any physical theory about the structure of objects. We can see this through the application of AP to both physically atomistic and 'gunky' (devoid of atoms) possible worlds. It is in fact consistent with both: the atoms making up a composite object in the first world may exist independently of each other, as may masses of gunk which are parts of a larger mass in the second.

A well-worn example also illustrates the doctrine nicely: according to AP, Michelangelo did not actually create his *David* from a block of marble. It was already there, along with many other overlapping shapes; his skill lay in bringing out that particular shape rather than one of the other less interesting ones. As such AP has much in common with what van Inwagen calls the Doctrine of Arbitrary Undetached Parts (DAUP):

> For every material object M, if R is the region of space occupied by M at time *t*, and if sub-R is *any* occupiable sub-region of R *whatever*, there exists a material object that occupies the region sub-R at *t*
>
> van Inwagen 2001, p.75

I am not sure whether AP and DAUP should be treated as different formulations of the same doctrine – they are certainly closely similar, but the move from decomposition in AP to occupiable sub-regions in DAUP may well be significant (to be pedantic, Holden phrases AP in terms of division rather than decomposition, but given we are talking primarily about spatial entities I think this minor obfuscation can pass). I intend to leave this question open, and say no more about DAUP, except that the challenge van Inwagen raises to the principle has, to my mind, been adequately deflected by Parsons (2004) who argues that his

[64] Strictly, the actual parts doctrine only claims that the parts which could exist independently according to logical possibility do in fact exist (Holden 2004, pp.12, 82), but we can pass over this detail for the time being.

challenge only works by rejecting the 'remainder principle', a theorem of classical mereology independent of DAUP which states that if *x* is not part of *y* then there is some thing produced by subtracting *y* from *x*.

Although perhaps ultimately acceptable, AP in itself takes no account of the strong conflicting intuition that Michelangelo did not just discover his *David* in a block of stone, but he *created it from* the block. On this line of thinking, it is simply wrong to say that the statue already existed as an independent entity before the artist's work was done. Likewise, the halves of the uncut apples in figure 5 do not exist over and above the apples themselves; they only come into full being at the moment of cutting. It does not follow from this that we cannot think or talk about the undetached halves at all, only that until the apple is cut the two halves are in some sense dependent upon the whole apple. For example it might be that while the apple can be physically divided into two halves, before that happens the halves only exist by virtue of a cognitive act of division. Again following Holden, we shall call this view the potential parts doctrine (PP), with different versions arising from different ways of cashing out this dependence principle.[65] Although rather less popular than AP it has its roots in Aristotle's hylomorphic distinction between matter and form, and is supported in Hobbes' *De Corpore* and Kant's *Critique of Pure Reason* (for other PP supporters see Holden ch.2). In the contemporary literature it is not wildly popular (although it is supported by van Inwagen to the extent that his dismissal of DAUP also rejects AP), though as we shall see Casati & Varzi (1999) take at least its motivations seriously.

Whether AP or PP ultimately turns out to be correct, I do think that some effort should be made to take into account the intuitions which fuel both positions.[66] On the one hand, Michelangelo did not cause any new matter to come into existence by chipping away at a block of marble. On the other, the matter which makes up the statue seems to have different properties when still part of the block than after the statue is carved. We have seen that the opposition between these two doctrines comes to light in situations where

[65] Holden himself classifies different varieties according to how they use four principles of divisibility: metaphysical, formal, physical and intellectual. Different combinations of these will also produce different dependencies between parts and wholes, so I take this formulation to be compatible with Holden's analysis.
[66] As mentioned at the start of this section, for current purposes we are assuming there are undetached parts in *some* sense to be established. Mereological nihilism, for example, bypasses the AP versus PP debate by denying that there are any parts, but we have already disregarded that view.

the relevant boundaries do not follow physical discontinuities in matter – as in the half of an uncut apple, or the as-yet unrealised *David*. Recall that Smith & Varzi's (2000) distinction between *fiat* and *bona fide* boundaries also turns on similar examples; they talk about a mountain, a (coastal) bay and the stem of a champagne glass (p.404), amongst other examples, but the point is the same in each case – *bona fide* boundaries always follow physical discontinuities, whereas *fiat* boundaries may or may not do so. This is no coincidence – can the *fiat* versus *bona fide* distinction be used to settle the question posed in our third problem case? Sadly not, but it is nevertheless a valuable conceptual resource in general, one well worth exploring for its own sake as well as to see why it falls short as an analysis of undivided parts.

4.2.4 Fiat versus Bona Fide Boundaries

Geographical boundaries are frequently set according to some sort of physical divide or discontinuity, such as a cliff, mountain range or river. Imagine two territories *Left* and *Right* are separated by a river which divides them. Both territories end at the river's nearside bank, with the river itself being neutral territory. Over time the river meanders, raising the question of whether the boundaries have remained in place or have themselves moved:

Figure 7: Some possible locations of the boundaries between *Left* and *Right*, before and after river meandering.

In a case like this it seems likely that most would agree the boundaries have moved as well as in the picture on the far right in figure 6 (if in doubt, compare this case with coastal erosion), but we should not expect this to be a generally applicable principle. If we accept that the boundaries have moved, we are still presented with a further problem, which has a much less obvious solution. As the river continues to meander it will eventually double up

on itself to form a small island, as in the leftmost picture in figure 7. If the boundaries of *Left* and *Right* still follow the river banks then the island forms a third territory separate from both: *Middle*. There is no clear criterion for deciding which of the original sides it should belong to – if indeed it belongs to either.

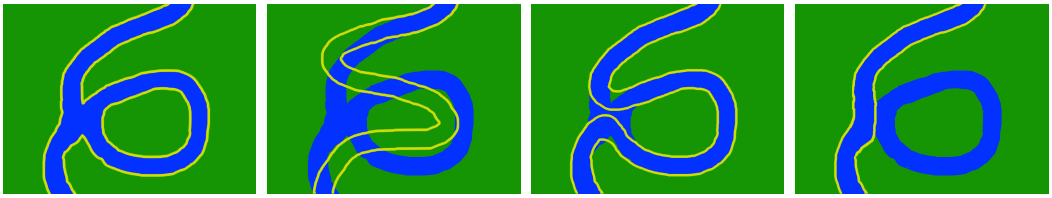

Figure 8: Four possible sets of boundaries, after further river meandering.

Various other options are of course possible as well, corresponding to the second, third and fourth pictures in figure 7: we could insist that the original boundaries stay in place no matter what (remembering where they are might be tricky, but that's another matter), the boundaries from before the island was formed stay in place, or that the island and ox-bow lake surrounding it are bypassed and the boundaries follow the most direct flow of the river. Others still are possible, but I think these are enough to make the two related points which should be drawn from the example.[67]

The first point is that the *fiat* versus *bona fide* distinction for boundaries, or one like it, is very difficult if not impossible to avoid. As soon as the river begins to meander it is clear that a decision has to be made regarding the status of the boundaries – must they follow the physical discontinuities of the river banks, or are its locations independent of (though originally derived from) them? Even if there were a clear, systematic answer to this question, the fact of its being raised is enough to show the distinction is worth taking seriously. Plausibly, the example above demonstrates that whether geographic boundaries such as these should follow physical discontinuities ought to be assessed on a case-by-case basis. When limited meandering has occurred, as in figure 6, it seems very sensible for the boundaries to move as well. Once the island forms this may well change, with either the

[67] There are of course many other interesting points regarding boundaries to be drawn from this example and others. Varzi (2004) in particular discusses several pertinent examples and a range of theories which address them.

third or the fourth picture in figure 7 depicting a more attractive solution. Most likely, whichever one is chosen will come down to specific factors – what if the river is too wide to cross, or the island too boggy to be used for anything? In any case, a decision will have to be made concerning whether the boundaries invariably follow the river banks or not.[68]

The second point is that while formulating the *fiat* versus *bona fide* distinction involves quite sophisticated spatial concepts, our river bank example which it addresses is really very simple. Far from arcane thought experiments located in distant possible worlds, or abstruse geometric or algebraic demonstrations, all the example needs to set up is a river and the concepts of boundary and territory.[69] As such the river bank example is readily expressible in lay terms, and what's more would probably have been comprehensible to even our very distant ancestors. Just how sophisticated the *fiat* versus *bona fide* distinction is compared to this is worth dwelling on for a few more moments. Smith & Varzi (2000) set up the distinction in terms of the opposition between Bolzano's and Brentano's theories of spatial boundaries. Bolzano's theory states that whenever two things are in contact there is a single boundary which belongs to one and not the other (in topological terms one includes its closure, so is 'closed', and the other does not, so is 'open'), whereas Brentano dramatically claimed this was 'monstrous' and argued that both things possess boundaries which depend upon those things to exist, and which overlap in cases of contact.

Smith & Varzi hybridise these views by taking *fiat* boundaries to be Brentanian and *bona fide* ones to be Bolzanian, arguing that the distinction cannot be properly made in a purely Bolzanian framework. I take it that the argument is sound, in which case the distinction between these types of boundary requires the rather sophisticated concept of container space. Container space is space conceived as an empty background against which objects are located and move around. Historically it was a modern development, advocated by Descartes and Newton, and it persists to some extent in substantivalist views of space-time

[68] Note this isn't quite the same decision as whether the boundaries are *fiat* or not. If they don't follow the river banks they are definitely *fiat*, if they do follow the banks then they could be either.

[69] Of course I'm assuming here various other basic concepts as well - time, space, matter, change and so on – but I take it there's no need to spell them all out. To be honest I don't think there's much prospect of doing so in the foreseeable future, as it would amount to a fairly comprehensive descriptive ontology for everyday, or 'commonsense', ways of thinking. As far as I can see that goal remains a long way off. This thesis is much less ambitious in its overall scope, in seeking to prune one term – 'representation' – from philosophers' own highly theoretical ontologies.

motivated by the General Theory of Relativity.[70] It needs to be in place to contrast Brentanian and Bolzanian boundaries, since they agree on all features of cases of contact between spatial entities except the number of boundaries involved and their locations; space itself is no different whichever theory we accept, so serves as a neutral background to the entities within it. Scientific concepts do filter down to the lay masses over time, but I'm inclined to think that this particular one has not yet. Certainly the *fiat* versus *bona fide* distinction would have been incomprehensible to our very distant ancestors, as they did not possess the conceptual resources to make sense of it.

This is a perfect demonstration of the tripartite nature of ontology, and another vindication of the value of viewing ontology in this way. We encounter an interesting phenomenon – boundaries – and seek to extract *prima facie* information about it. Empirical examples show that boundaries may be understood as dependent upon or separate from physical discontinuities, and by distilling and codifying this fact Smith and Varzi produce the formal theory of *fiat* and *bona fide* boundaries. They then formalise the theory, by rendering it in a logical form. So far, the theory serves well to deal with an otherwise puzzling issue.

One aspect of boundaries left out by Smith & Varzi's distinction is that they may well not be particularly homogeneous. A significant proportion of their examples involve rivers and oceans, but as Steinberg (2001) convincingly argues there are many different ways of conceiving and managing ocean space – and by extension I take this to apply equally to inland waters, and to spaces in general. He frames his discussion around a naval incident in 1990 when twenty-one containers, including five of Nike goods, were lost overboard near Alaska.

> Four of the five Nike containers opened, and 61,280 shoes began a long journey eastward to the coast of North America. Over the next two years, more than 1,600 Nikes were recovered on the beaches of British Columbia, Washington and Oregon.
>
> Steinberg 2001, p.1

[70] See e.g. Hoefer (1996) for a sophisticated survey and analysis of versions of space-time substantivalism.

Not surprisingly, enterprising local residents collected the shoes and some sold them on, going to considerable lengths to get hold of matching pairs. To them the ocean was a provider of goods, whereas to Nike it represented a featureless distance to be crossed, and the insurance company viewed the ocean as a space of discrete places and events (*ibid.* pp.2-3). Along with these different conceptions of the ocean we might well hypothesise different types of boundary. From Nike's perspective the ocean is a purely spatial entity to be crossed, so has boundaries which are straightforwardly spatial, whereas as far as the shoe collectors were concerned most of the spatial features of the ocean were irrelevant – their interest lay in its depositing shoes on the shore. As such we might take it that they understood the ocean's coastal boundaries in terms of fulfilling this function rather than its spatial characteristics. It seems plausible that the insurer's view of the ocean's internal and external boundaries would make use of a combination of spatial and functional features.

This suggests that just because obvious examples of boundaries are of ones which lie between spatial entities (all of Smith & Varzi's examples are such), it does not necessarily follow that boundaries themselves should be characterised in exclusively spatial terms. The *fiat* versus *bona fide* distinction is inherently spatial, and while we have seen that there is a need for this distinction or something like it, there is no reason I can see to think that a fuller account of boundaries will be characterised in exclusively, or even primarily, spatial terms.

The Nike incident provides some *prima facie* evidence to suggest that spaces, and hence boundaries, are at times conceived in partially or entirely functional terms. It might be that function (perhaps construed according to one of the options canvassed earlier) provides an ultimately more suitable characterisation of boundaries than space, although in the absence of a fully-fledged theory we are pointing out an alternative line of inquiry rather than arguing in its favour. Incidentally, for better or worse, this view of boundaries would be much less revisionary in the Strawsonian sense of going beyond everyday understanding, since functions are easily recognisable without similarly sophisticated conceptual distinctions in place. After all, even young infants are capable of recognising the functions of various kinds of objects – to indulge in anecdotal evidence, my eighteen month old daughters know very well that large metal keys are for opening doors, small keys are for

windows, and large black keys are for starting cars. Alternatively, the considerations thrown up by the Nike case may be germane to the *fiat* versus *bona fide* distinction, by illustrating the criteria for *fiat* boundaries in the cases of the local residents and the insurer; different needs or interests may well create different *fiat* boundaries, but the boundaries themselves remain spatial entities. I'm not sure which of these options is preferable – much less which, if any, is the right one – so I intend to reserve judgment for the time being.

But what does all this tell us about whether and how the *fiat* versus *bona fide* distinction can answer the question in our third problem case? We have seen from the river example that the distinction, or one like it, is inevitable given changes in the environmental features which boundaries follow, and so that the distinction is a valuable resource in providing a philosophical theory of boundaries in general. Considerations arising from the Nike example suggest that a full theory of boundaries lies yet some distance in the future. However, the distinction cannot in itself provide any genuinely informative analysis of potential parts. The biconditional '*x* is a potential part iff it possesses a *fiat* boundary' is inadequate as an analysis because some or all whole objects may possess *fiat* boundaries. We can add some sort of supplementary condition *θ* to give '*x* is a potential part iff it possesses a *fiat* boundary *& θ*', but this is scarcely better. We characterised potential parts as those whose boundaries do not follow a physical discontinuity in the whole, such as in our half an apple, part of a homogeneous mass of gunk, or the as-yet uncarved block of marble. *Fiat* boundaries have been characterised in two ways. First, as those which need not follow physical discontinuities, and second as being Brentanian, i.e. depending upon the things they bound and being capable of overlap. The second simply comes from Smith & Varzi's analysis of the first, and using the first renders the proposed analysis circular. Both the left and right hand sides of the biconditional are defined in terms of physical discontinuities – particularly so if we adopt the obvious candidate for *θ*, which is that the fiat boundary does not in fact follow a physical discontinuity. I think it would be unfair to say that the biconditional is not informative at all, but it falls far short of a genuine analysis of potential parthood.

4.3 Countermereotoplogy

So what to make of the intuition behind the potential parts doctrine? Casati & Varzi (1999) aim to account for the intuition that parts such as the halves of an uncut apple are potential parts of the whole – and so are dependent upon it – whilst remaining true to the letter of the actual parts doctrine, which takes the parts to be independent. If successful this will allow the AP explanation of examples such as our third problem case to stand, while explaining away the contrasting intuitions which motivate PP.

Casati & Varzi's strategy is to propose a modified form of Lewis' counterpart theory, where the two halves of the uncut apple are entities existing in a different possible world where the apple has in fact been cut. They call this theory countermereotopology.[71] Spatial objects such as apples are, according to the theory, trans-world individuals: the whole apple exists in one possible world, whereas its two halves exist in some other possible world where the apple has already been cut. The two halves are potential parts of the apple in the sense that they do not actually exist here and now, but they are also actual objects – though not actually parts of the whole apple – in the world in which they do exist. Thus, Casati & Varzi maintain, we can preserve the intuition that undetached parts such as the apple halves in our third problem case are potential objects, whilst nevertheless treating them as actual ones. In this section we will see that countermereotopology is not tenable, as shown by an adapted version of Cresswell's (2004) arguments against Lewis' counterpart theory. First we will outline Lewis' theory and Cresswell's original arguments, then adapt them to our purposes.

[71] Casati & Varzi describe their motivation slightly differently, as wanting to take account of the modality of the potential parts doctrine without abandoning mereological extensionality (see pp.100-3), but both their descriptions and mine plausibly amount to the same thing.

4.3.1 Lewis' Counterpart Theory

Counterpart theory enables talk of possibility and necessity within ordinary first-order logic by acting as a substitute for identity – when I say, for example, that I could have been a contender this is understood not as a modal fact about me as I am, but as an actual fact about someone in another possible world who is suitably similar to me. The contender is my counterpart (although I may not be his), and the truth of my claim is grounded in his existence in some possible world plus his being my counterpart as opposed to someone else's or nobody's at all. The counterpart relation, then is a similarity relation: it is originally presented by Lewis (1968) as being reflexive, asymmetric and intransitive (pp.28-9, referenced in his 1986).

The theory itself is composed of four primitives, eight postulates, and a number of informal comments. The primitives represent the concepts of being a possible world, being in a specified possible world, being actual and being a counterpart of some specified thing. The postulates are, informally:

1. Nothing is in anything except a world.
2. Nothing is in two worlds.
3. Whatever is a counterpart is in a world.
4. Whatever has a counterpart is in a world.
5. Nothing is a counterpart of anything else in its world.
6. Anything in a world is a counterpart of itself.
7. Some world contains all and only actual things.
8. Something is actual.

Also, an object can have more than one counterpart in a world, and can be counterpart to more than one object in a world; an object need not have a counterpart in every world, nor must any object in one world be a counterpart of something in another. A further stipulation is that within a world an object is its only counterpart, but Lewis (1986) says he would be prepared to drop this in some cases (see also Hazen 1979, Cresswell 2004 for

arguments against the stipulation). Lewis (1968) also provides a translation schema from quantified modal logic into counterpart-theoretic terms, but that is distinct from the theory proper so we need not dwell on it here (see e.g. Hunter & Seager 1981 for further defence of this claim).

Casati and Varzi adapt this theory by adding to their mereotopological apparatus the idea of a partial counterpart, with the partial counterpart relation being asymmetric and intransitive like Lewis' relation, but irreflexive. The two halves of an apple are *partial* counterparts of the whole apple, in the sense that there is a possible world in which they are themselves wholes – such as the world in which the apple is cut in half. Undetached parts of an object exist in a different possible world (or worlds) to that object itself, and only in worlds where there is no counterpart to the original object (Casati & Varzi 1999 p.103). Since Lewis' counterpart theory, and Casati & Varzi's revisions, can be expressed in first-order logic plus identity, this allows the employment of modal concepts without abandoning classical mereology or topology:

> [T]he mereotopology stays, and to treat parts as potential is simply to treat them as belonging to different possible worlds than the wholes to which they are mereologically attached
>
> Casati &Varzi 1999, p.103

Likewise, it takes account of the intuition that undetached parts are *potential* objects whilst treating them as actual objects, albeit ones which exist in different possible worlds.

Although popular, Lewis' counterpart theory is still controversial, with arguments against it falling into four broad types. The first type, exemplified in Kripke's (1972) *Naming and Necessity* and Plantinga's (1974) *The Nature of Necessity*, focus upon the apparent counterintuitiveness of Lewis' theory. These arguments claim that when trying to express a possibility about an object in one world counterpart theory is in fact talking about some other object in another world, and that this is a mistake. When talking about something, we are talking about *that thing* and nothing else. But as Hazen (1979 p.323) observes, this line of argument confuses the semantics of Lewis' theory with those of the natural language expressions it is an analysis of, and our ordinary intuitions apply to the latter rather than the

former. It is almost certainly too strong to say that an analysis of ordinary modal discourse can be as counterintuitive as we like, but the fact that such an analysis does run counter to ordinary intuitions is a relatively weak argument against it. Arguments of the second type are stronger in the sense of being more persuasive, but they are concerned with the translation schema between counterpart theory and quantified modal logic, and so are limited to being mildly revisionary (see Lewis' original 1968 schema, revisions by Forbes 1982, 1990, and Ramachandran 1989, 1990a, 1990b, and also Hunter & Seager for the denial that any schema is needed). Whatever translation schema is adopted will not have any substantial impact on the theory itself, so in turn these arguments will not have any impact on Casati and Varzi's partial counterpart relation. The third variety are technical objections, although these are relatively scarce: much of the attraction of counterpart theory, I believe, comes from its technical sophistication and elegance. That is not to say that none are worth taking seriously. Fara's (2007) criticisms may well be damning, despite Melia's (2007) lukewarm response.

The fourth variety of argument against Lewis' theory is the semantic type, those concerning the intended meaning of the theory and its terms. Such arguments are perhaps the least well-represented in the literature, with the leading example being Cresswell's (2004) argument that the counterpart relation is in at least some cases symmetric and transitive – and furthermore that it is in some cases both, i.e. it is an equivalence relation. Although it is doubtful whether the argument holds against Lewis' theory, when applied to Casati and Varzi's partial counterpart relation, it in fact shows that in most or all cases partial counterparthood is transitive. This formal inconsistency with Lewisian counterpart theory illustrates that there is a fundamental difference between parthood and counterparthood which cannot be resolved. Therefore countermereotopology fails as an account of undetached parts (whether potential or actual, neither or both) such as in our third problem case.

4.3.2 Cresswell's Semantic Arguments

Cresswell argues that in at least some cases the counterpart relation in symmetric, and that given symmetry it is in at least some of those cases it is also transitive. Thus the relation is in at least some cases an equivalence relation. The first two arguments are presented in terms of a temporal counterpart relation analogous to Lewis' relation, then applied them to the modal case by analogy; both also turn on the fact that symmetry and transitivity are concerned with what happens across at least two iterations of the counterpart relation (p.31). The argument for symmetry takes the form of a *reductio*, and runs as follows. Assuming asymmetry of temporal counterparts, the following is possible:

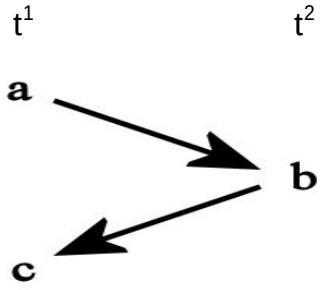

Figure 9: Asymmetric temporal counterparthood.

Here times are intended to be analogous to Lewisian possible worlds, and so all exist equally; the arrows indicate that *b* is the future counterpart of *a* whereas *c* is the past counterpart of *b*. According to Cresswell this falsifies the intuitively valid inference:

(α) If I am now bald then I will always have once been bald.

Were the inference invalid then there would have to be a tenable interpretation of (α) such that at some time it would be true that I am bald, but false for at least one later time that I had once been bald. I think it is clear we should agree with Cresswell that this is highly implausible, and so conclude that the temporal counterpart relation is in this case

symmetric. If the analogy with the modal relation holds, then some modal counterpart relations are also symmetric.[72]

The argument for transitivity runs as follows. Assuming transitivity, the following is possible:

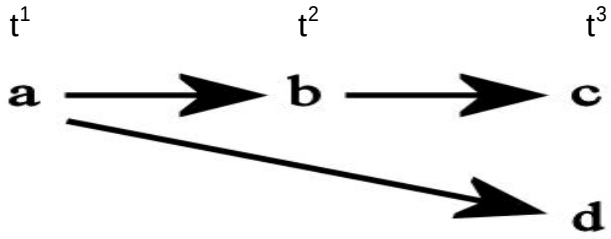

Figure 10: Transitive temporal counterparthood.

Here *b* is the counterpart of *a* at the later time t^2 and *c* is the counterpart of *b* at the later t^3, but *d* is the counterpart of *a* at t^3. According to Cresswell if *a*, *b*, and *d* are bald but *c* is not – i.e. if the counterpart relation is intransitive – then the following inference is falsified:

(β) If today I'm always going to be bald then tomorrow I'm always going to be bald.

As this inference is intuitively both valid and sound, the counterpart relation therefore is in this case transitive.

4.3.3 A Temporal/Modal Analogy

I take it that both of these arguments are themselves persuasive, showing that at least some *temporal* counterpart relations must be symmetric and some transitive (for Cresswell's claim that some are both see p.33). However, they cannot be straightforwardly applied to the modal case. In the above arguments it has been tacitly assumed that:

[72] Given Lewis' 1986 admission that some counterpart relations may be symmetric after all, this is a minor victory. Cresswell presses it further by observing that Lewis has failed to show that any reasonable relations are not symmetric.

1) We have a definite point of reference – the present – by which we can measure transitions between different times.
2) A plurality of times always forms a linearly ordered sequence (e.g. past, present, future; yesterday, today, tomorrow; earlier, now, later).

Both of these assumptions are, I shall take it, reasonable enough for a temporal relation, but in the modal case while an analogue of 1) is acceptable, a modal version of 2) is straightforwardly false.[73]

Cresswell recognises the first of these assumptions, but not the second, and to preserve a temporal-modal analogy suggests that in the modal case we should include an embedded operator which has the role of bringing us back to the world we started from (p.32). This is sufficient to apply the above argument for symmetry to modality as (α) only involves the present as t^1 and whatever later time we take to be t^2; a modal analogue would likewise only involve two worlds, one being the actual world, and so the operator would range over all transitions between all worlds considered in this example. Assumption 2) is much easier to miss, since in example (α) it effectively collapses into 1). It can be seen to be in place, however, if we unpack (α) into an equivalent:

(α*) If F*x* at present time t^1 then for all future time t^2 there exists a past time t^0 such that F*x* at t^0.

In (α) t^1 and t^0 are the same time, so there 2) is unproblematic. As a result Cresswell's argument that at least *some* modal counterpart relations are symmetric holds, though we should be cautious about assuming it shows that many or all counterpart relations are.

Forming a modal analogue of the transitivity argument is more complex, as it must involve three possible worlds (at most one of which may be the actual world) and so the two assumptions can be more easily seen to be distinct. We have seen that transitivity occurs across two or more iterations of a relation, and hence between at least three relata. Where these relata are times, at least two will be non-present times, and *if* all we had to go on

[73] Since this is in no way a dissertation on the nature of temporality I make no attempt to defend this sweeping claim about time; I hope it is reasonable enough at first blush to be taken for granted here, or at least excused.

were a list of times and knowledge that a specific one of them were the present, we would not be able to say of two non-present times which were the earlier and which the later. Not being able to put times into any sort of order apart from 'now' versus 'not now' would be problematic: it would block Cresswell's argument for transitivity as we would have no way of telling whether t^3 was later than t^1. If t^3 is earlier than t^1 then there is no problem for transitivity – after all, hirsuteness in the past is no guide, or obstacle, to baldness in the future. If we cannot tell which time is earlier then we cannot tell if there is a problem or not, which is rarely a good thing.

Thankfully the presence of 2) means we do not have to worry about this – times are naturally ordered themselves. The problem is that in order for the argument from transitivity of the temporal counterpart relation to carry over to the modal case, that is for intransitivity of the modal counterpart relation to falsify a suitable analogue of (β), there must likewise be a modal equivalent of 2), to the effect that possible worlds form a linear ordered sequence. It might be thought, as Cresswell claims (p.33), that an actuality operator plus symmetry is sufficient to make the analogy, but I fail to see how this could be right: one fixed point does not a sequence make. The only way I can see to make Cresswell's transitivity argument work would be to adopt a modal equivalent of 2), to the effect that possible worlds form a linear ordered sequence – effectively, a dimension – but this is highly dubious. Just this lack of any natural ordering is a key point of disanalogy between times and possible worlds! As Quine put it,

> The devastating difference is that the series of momentary cross sections of our real world is uniquely imposed on us, for better or for worse, whereas all manner of paths of continuous gradation from one possible world to another are free for the thinking up
>
> 1981, p.127

Even though it is perfectly possible to think up a relation between possible worlds which does order them into a linear sequence, such as their overall similarity to the actual world, the point is that there is 'no *modal* comparative relation between these worlds which *independently* puts them into any linear order whatsoever' (Lowe 1986, p.197). Times can be put into a linear sequence according to the purely temporal relations which hold

between them (say, of earlier and later), whereas no plausible candidate for a purely modal equivalent seems to be forthcoming.[74] In its absence, an analogy between time and modality which draws upon the sequential nature of times must surely fail. The strategy of tackling tricky modal issues through simpler temporal ones which Cresswell adopts (see also his 2005, p.435) is not necessarily wrongheaded for every situation, but it needs to be pursued with care to the differences between the two.

So Cresswell's arguments for symmetry and transitivity of the Lewisian counterpart relation meet with mixed success. I take it that the modal argument for symmetry is strong enough to establish that in at least some cases Lewis' counterpart relation is symmetric, while the modal argument for transitivity fails. Applied to Casati and Varzi's partial counterpart relation the converse holds true: adapted versions of Cresswell's arguments fail to demonstrate symmetry but they do show the relation to be transitive in general. A further argument by Simons demonstrates that the partial counterpart relation is symmetric in at least some cases. This demonstrates an irreconcilable difference between counterparts on the one hand, and partial counterparts on the other.

4.3.4 Counterparts and Partial Counterparts

Let's start with the transitivity argument first, as our discussion here will help inform the symmetry argument next. We have seen that Cresswell's argument fails in the modal case because, unlike with times, there is no independent ordering of possible worlds beyond those we arbitrarily assign to them. In contrast parts, like times, are intrinsically ordered. This is trivially true in virtue of the mereological formalisation of parthood as a partial ordering. Even were we to have rejected mereology outright – which we have not – it is still very plausible as an analytic truth about parthood that parts and wholes intrinsically form a linear hierarchy. *Reductio* arguments to support this claim shouldn't be too hard to formulate, though I really think none are needed.

[74] I take it in agreement with Lowe that modal accessibility is a fictional relation, and so cannot be said to hold of possible worlds independently of the logical systems they are invoked in.

So it quickly follows that there is in fact a close analogy between the temporal and partial counterpart relations in exactly the way in which Cresswell's analogy between the temporal and modal counterpart relations fails: both temporal and partial counterpart relations are intrinsically ordered. We might remember Lowe's objection that there is still no *modal* comparative relation between worlds to order them. However, to think Lowe's point applies here as well is to partially miss the point of countermereotopology; it isn't a theory of modality, it's a theory of parthood and there is a *mereological* comparative relation between the parts and wholes which exist in different possible worlds.[75]

Adapting Cresswell's transitivity argument to exploit an analogy between time and parthood succeeds in showing that countermereotopological relations must be in general transitive. Against Casati & Varzi, it cannot be intransitive in general but is either non-transitive or transitive. In mundane spatial cases it seems that it will invariably be transitive, as it is not too difficult to think up suitable analogues to Cresswell's (β). Take our third problem case, of half an apple, and let's assume it's the left half relative to some plane for simplicity. This is equivalent to the examples Casati and Varzi use themselves. If the left half of an apple is an undetached part (or partial counterpart) of the whole, whose left half in turn is an undetached part of it, then of course the result of this second potential halving is an undetached part of the whole apple – it's nothing other than the left quarter of the apple! It seems bizarre to think that the parthood relation here might not be transitive, and the same point holds equally well if we swap 'apple' and 'left half' for any other material and spatial concepts respectively. So countermereotopology is in general transitive.[76]

What about adapting Cresswell's symmetry argument? This fares less well; as we have seen the transitivity argument works against countermereotopology precisely because parts form a linearly ordered sequence, thus satisfying Cresswell's second implicit assumption which requires this to be the case. In his symmetry argument this assumption of an ordered sequencing amongst the relata collapses into the assumption that there is a definite point of

[75] To be clear, Lowe himself does not make this mistake.

[76] If countermereotopology were extended to non-spatial objects, or possibly non-standard geometries, then the argument for transitivity would be proportionately weakened. Casati and Varzi explicitly do not make this extension, and to the best of my knowledge no-one else has either. There may also be some ordinary spatial cases where intransitivity of the relation holds, but I doubt it.

reference (for the modal case, the actual world), and as a result his argument goes through for at least some modal counterparts. Generating comparable examples to Cresswell's (α), only for the partial counterpart relation, is a tricky business. Instead of times we would have parts across possible worlds, and for the convenience of using an irreflexive parthood relation we would need to multiply the entities involved to at least two non-identical ones. The reference to baldness should be dropped (or replaced by something with a spatial character). The result would be something like this:

> (α^p) If X is a potential part of Y, then X is a potential part of Y regardless of which of a series of possible worlds X is located in.

Unfortunately, this is not very helpful – because, again, possible worlds are not intrinsically ordered so talk of series of possible worlds has no value in this context. So is the partial counterpart relation symmetric or not? I do think a bit of intuition can guide us here: it would be frankly bizarre if the relation were symmetric in general. Think again of the apple case. If the left half of the apple is a partial counterpart (undetached part) of the whole apple, and this relation is symmetric, then the whole apple is a partial counterpart of its left half. Recall that we characterised this relation in terms of a possible cutting or division which has not actually occurred; there is simply no way to cut up half an apple to make a whole one![77] We have already seen that this example generalises well to other spatial cases, so the partial counterpart relation is in many or most cases asymmetric (I take it anti-symmetry is quite implausible).

Simons has suggested an ingenious example (in conversation) to demonstrate that cross-world mereological relations may be symmetrical in some cases. In the actual world w0, B is an accurate statue of Cromwell complete with warts while A is the mass of marble representing the warts. A is a proper part of B. In world w1 the counterpart of Cromwell has pockmarks instead of warts. In w1 B – the accurate statue of Cromwell's counterpart-in-w1 – also has pockmarks. If the pockmarks were filled, this would produce a mass of marble the

[77] Admittedly the Biblical story of the feeding of the 5,000 does suggest that there may be a way to break up bread and fish to produce greater quantities, which could be interpreted as turning parts into wholes. However, this need not have been a mereological miracle – it could have been a case of spontaneous generation, for example – and the story's literal truth is not beyond doubt.

same size and shape as A in w0. Hence B-in-w1 is a proper part of A-in-w1. Cromwell-in-w0 is the counterpart of Cromwell-in-w1 and vice versa, likewise for B-in-w0 and B-in-w1, and for A-in-w0 and A-in-w1. From this it follows that in w0 A is a proper part of B, whereas in w1 B is a proper part of A.

We can see from examples that the partial counterpart relation cannot always be symmetric (by apples and their halves), yet it is symmetric in some cases (by the statues of Cromwells). From this it follows that partial counterparthood is non-symmetric, meaning it may be symmetric or asymmetric depending on context.

4.3.5 Undetached Parts, Counterparts and Partial Counterparts

The analogy with Lewis' counterpart theory has become strained to breaking point. Recall that the modal counterpart relation is reflexive, either non-symmetric or symmetric depending on how we interpret Cresswell's argument (but remember Lewis' 1986 admission of symmetry in some cases), and intransitive. In contrast, the partial counterpart relation is irreflexive, non-symmetric and transitive. The two relations have very different properties; they are like chalk and cheese. This undermines Casati and Varzi's interpretation of their formal theory – the parthood relation between wholes and undetached parts is utterly unlike the modal counterpart relation, so adapting Lewis' theory to apply to parthood is plain inappropriate. It is a case of well-meaningly shoehorning a square peg into a round hole.

This is the real problem: counterpart theory and the mereology of undetached parts are very different sorts of theories. There is a line of thinking which makes putting the two together attractive; after all the counterpart relation functions as a substitute for identity, and according to classical mereology identity is a limit case of parthood (according to Lewis and Sider amongst others, the connection between the two may be more intimate still). What's more, counterpart theory is a theory of modality, and undetached parts are parts which *could* be separated, but haven't been. Counterpart theory even preserves

extensionality, so we can keep mereotopology as it stands – consistent with the actual parts doctrine – whilst taking into account the modality central to the potential parts doctrine. Surely the two are made for each other? The reason why they are not is this: counterpart theory tells us about what *could happen to something, but hasn't*, in terms of another thing which is a stand-in for our original. What cases of undetached parts require is a theory which tells us about the composition of an object *as it is now*, with the AP vs. PP debate turning on whether this needs to be cashed out in terms of what could happen to it. Chalk and cheese. The only useful way counterpart theory could be applied to undetached parts is if PP is correct and we require an account of modality to supplement an already in-place mereological theory. Whether even this would be worthwhile is questionable.

4.3.6 Undetached Parts, Again

So what are we left with? Countermereotopology fails as a substantive account of undetached parts because all it amounts to is a list of properties for the relation plus a highly tenuous analogy with counterpart theory. It is not capable of explaining the intuitions which underlie PP in a theoretical framework which supports AP. Stipulating some plausible formal constraints is a very different activity from actually providing a theory which satisfies them *and* explains the relevant phenomena. That is not to say countermereotopology is entirely without merit. Suggesting plausible constraints on the relation which hold in cases of undetached parthood is a valuable activity, and could well form the basis of a more plausible theory – one without the misguided association with counterpart theory.

Smith and Varzi's *fiat* versus *bona fide* distinction likewise fails as an analysis of potential parts (though we should be clear that was not the authors' intention for it). The solution to our third problem case provided by mereotopology is acceptable, but only acceptable to someone who endorses the actual parts doctrine. Here again we see descriptive ontology constraining the choice of possible formal theories.

The potential parts doctrine can equally provide a schematic explanation of our third problem case: an uncut apple half is a part of the apple which shares all its parts with the whole apple (though not vice-versa). A physical act of cutting would divide the apple into two detached halves; prior to that the division between the undetached apple halves is not physical (it may be – but is not necessarily – cognitive). Likewise all the parts of Michelangelo's *David* exist as parts of the uncut marble, but only through its carving is the statue produced as a *separate* object. There are numerous ways in which this schematic answer could be fleshed out; in the following sections we will use Kit Fine's theory of rigid and variable embodiment to illustrate how it can be done. We shall see that this PP-style theory is able to successfully account for all of our problem cases and provides a viable alternative to classical mereology. But first we shall question the eighth and final tenet of classical mereology: univocality. Drawing upon arguments made by Fine we shall see that it is false.

4.4 Is Parthood Univocal?

In this final section on mereology we shall see that its claimed univocality is false, on the strength of arguments by Fine (1999). We will also see that Fine's rival theory of rigid and variable embodiments are equally able to explain our three problem cases, but do not fall foul of the problems he raises for mereology, and can meet recent criticisms made by Koslicki (2007). Ultimately the issue of whether to adopt an AP-style theory such as classical mereology or a PP-style theory such as Fine's turns on what descriptive characterisation of things and their parts we prefer to choose.[78]

[78] Given the preceding discussion of other mereological principles I shall rule out the possibility that parthood is univocal but is not modelled by classical mereology. To the best of my knowledge there is no serious attempt to realise this option in the literature, and it would be very hard to rationalise with most non-mereological theories of parthood for reasons we will come to shortly.

What reason do we have to doubt that parthood is univocal? One strong motivation comes from philosophers who endorse an ontology of states of affairs, as it seems that they must be composed structurally. Armstrong takes this view, and argues for the supplementation of mereology with a distinct structured parthood relation:

> Lewis holds that mereological composition is the only form of composition there is... The moral drawn in the present work is that there has to be at least one other form of composition in the world. We have already found nonmereological composition in states of affairs.
>
> 1997, p.187

States of affairs are, however, sufficiently controversial that I shall put aside arguments for compositional plurality which depend upon their existence. A comparable point could be made in terms of, say structured universals – perhaps no less controversial – or indeed anything with structure (mental representations could be another case in point). However, given our discussion of extensionality and rearrangement of parts in chapter three where we saw that claims about structured and non-structured parthood are translatable *salva veritate*, we should be sceptical about appeals to structure forming the basis of a call to reject classical mereology. Armstrong is right to say that there must be at least one form of non-mereological composition (if not more!), but for the wrong reason.

To return to our opposition of AP-style and PP-style theories (where classical mereology is the principal modern example of the former), potential parts theories will in general require multiple parthood relations – tensed and tenseless ones.[79] Think of the left half of our apple – it is only a (potential) part of the apple prior to the apple's being cut and so the parthood relation there needs to be relativised to a range of times. On the other hand, PP theorists are likely to want to retain cases of parthood which are not relative to times, such as essential parts without which the whole would not exist, or would be significantly changed. But there is no necessary requirement to stop there – free from mereological univocality, the PP-style theorist is at liberty to posit a cornucopia of different parthood relations according to different sorts of composition, different domains, or both.

[79] I would like to thank John Hawthorne for helping me to see this point clearly (in conversation).

But is mereological uniqueness definitely false? Probably the strongest argument that it is comes from Kit Fine's 'Things and their Parts' (1999), which proposes two objections designed to show that classical mereology is fundamentally flawed in its characterisation of material objects – in brief, because it is temporally disjunctive. These are the so-called 'aggregative' and 'monster' objections. We will see that these objections are very powerful indeed, but given our qualified support for classical mereology so far we should not throw the baby out with the bath water. Rather, Fine's arguments show that mereology is restricted in scope to at most some classes of things and not material objects in general, and hence that mereological univocality rather than uniqueness is false. While the arguments certainly could be interpreted as disproving uniqueness, and need to be satisfactorily addressed to avoid this conclusion, the strength and popularity of a suitably modified classical mereology makes me reluctant to declare the issue to be settled one way or the other. Fine's arguments may yet be conclusive, but in the absence of any substantive attempt to reply on behalf of classical mereology, it is too soon to tell.

Having presented his arguments against classical mereology, Fine himself proposes a broadly PP-style neo-Aristotelian alternative to classical mereology which deals equally well with our three problem cases as supplemented classical yet is plurivocal.

4.4.1 The 'Aggregative' and 'Monster' Objections

> [On the classical mereological understanding of 'sum'], a sum of material things is regarded as being spread through time in much the same way as a material thing is ordinarily regarded as spread out in space. Thus the sum a + b + c + ... will exist *whenever* any of its components, a,b,c, ... , exists (just as it is located, at any time, *wherever* any of its components are located). It follows that under the proposed analysis of the ham sandwich, it will exist as soon as the piece of ham or either slice of bread exists. Yet surely this is not so. Surely the ham sandwich will not exist until the ham is actually placed between the two slices of bread. After all, one *makes* a ham sandwich; and to make something is to bring into existence something that formerly did not exist.
>
> Fine 1999, p.62

The problem here is that according to Fine classical mereology assigns the wrong spatial and temporal existence conditions to material objects (or at least to non-atomic ones). On the standard understanding taken by classical mereology, the mereological sum a + b + c will exist whenever and wherever any of a, b, or c exists in the same way as an arithmetic sum of three numbers will produce a non-zero result as long as one or more of the numbers is not zero. To avoid this would apparently require that mereology be made sensitive to the spatiotemporal location of all parts of the whole. It would also apparently need to be sensitive to their manner of arrangement as well: if ham and bread are at opposite ends of the room (or the Earth) then no whole they compose should count as a sandwich.

The problem seems clear. Advocates of classical mereology would of course agree that the ham and bread only make a sandwich when they are in a suitable arrangement – but insist that at those times when they are arranged sandwich-wise the mereological sum of the three just *is* the sandwich. Classical mereology models parthood *tout court*, parthood resulting in a sandwich would require an additional non-mereological restriction concerning arrangement in space and time. We have already seen that this popular strategy is fruitful, canvassing restrictions related to function, connection and several other criteria. The obvious candidate here is a temporal restriction, to only the times when there actually is a sandwich. Fine canvasses this option, observing that adding a temporal restriction on parthood serves to produce an 'extended' sense of part according to which,[80]

> [g]iven any two objects we may say that the first is, in this extended sense, a part of the second if the restriction of the first to the times at which the second exists is (in the unextended sense) a part of the second. Thus for the purpose of making judgments of part, we ignore what there is to the first object outside of the time at which the second exists.
>
> Fine 1999, p.63

It might be better to ignore the potentially confusing 'extended' parthood Fine introduces, and concentrate on the main point. Imposing a straightforward temporal restriction on

[80] I say 'extended' sense of part as I think this should more properly be considered to be a standard case of mereological parthood which incorporates additional restriction(s) in the same way as φ-restrictions discussed in chapter three. This has no impact on Fine's argument, however.

bread and ham to when they make a sandwich is equivalent to turning a blind eye to whatever the sandwich's parts make the rest of the time. This is bad news:

> Consider … the sum of the ham and all objects that existed only before or after the sandwich existed. Then the restriction of this sum to the time the sandwich exists is the same as the restriction of just the ham and hence must also be a part of the sandwich. But it is ludicrous to suppose that this monstrous object – of which … all merely past and future galaxies are parts – is itself a part of the ham sandwich.
>
> Fine 1999, p.63

Unrestricted composition guarantees the existence of a sum of the ham and all objects that existed only before and after the sandwich, and varying the temporal restriction on this sum will not alleviate the problem. All a different restriction will produce is a different monster, but a monster nonetheless.

For my part, I must admit I found both objections – and particularly the second – a little difficult to follow at first, and so for some time underestimated their value. In the interests of clarity, I think it is well worth examining their structure and content in more detail before evaluating their significance.[81] Both objections are best read as pressing implicit *reductio* arguments against the disjunctive account of parthood, with the monster argument incorporating a presumed reply to the aggregative argument. The first *reductio*, implicit in the aggregative objection, goes like this:

P1. Composition is disjunctive across space and time.
C1. Composite objects therefore exist whenever and wherever any one or more of their parts exist.

This is – says Fine, and I am inclined to agree – utterly at odds with what ordinary material objects are like, so either the premise P1 must be rejected or else the argument must be augmented by some additional premise to avoid the conclusion C1. The additional premise Fine supplies is a temporal restriction to times when all parts of the composite object exist and are suitably arranged:

[81] Koslicki (2007, pp.140-143) also provides a useful and somewhat different exegesis of the two objections, although I think the differences there are lie in manner of presentation rather than any substantive differences in how the objections are interpreted.

P2. Composite objects exist only when and where all of their parts exist at suitable times and places.

This is incorporated into the *reductio* presented by in the monster objection:

P1. Composition is disjunctive across space and time.
P2. Composite objects exist only when and where all of their parts exist at suitable times and places.
P3. If unrestricted composition is true, there will exist a 'monster' composed of one of the parts (e.g. ham) and other things which existed before and/or after the restricted time period (e.g. before and/or after the sandwich existed).
C1. No temporal restriction on disjunctive mereology will be able to differentiate between the genuine part (ham) and the 'monster' as parts of the composite object (sandwich).

This argument proceeds from abstract principles to specific cases, but its conclusion C1 is so outlandish that one or more of the premises must be false. We have already seen that unrestricted composition is highly defensible (and Fine does not challenge it). The temporal restriction in P2 is the only obvious way of meeting the reductio posed by the aggregative objection. Therefore, P1 must be false – composition is not *always* disjunctive across space and time (and, we can add to Fine's argument, since mereological composition is disjunctive is it follows that mereological univocality is false).

It should be clear that the combination of Fine's aggregative and monster objections provides a significant challenge to classical mereology as a fully adequate theory of parthood relations amongst material objects. Hence it is an objection to the claim that mereology is univocal. The aggregative objection demonstrates the need for a conjunctive account of parthood as well as the 'standard' disjunctive account incorporated into classical mereology (this theme is also explored in Fine 1994). The monster objection shows that there is no straightforward patch-up for a disjunctive theory to get around this problem.

The challenge the combined objections pose is also more significant than it might at first appear: on a casual reading it looks a lot like a problem about structure – a question of rearrangement of parts. The ham needs to be *between* the slices of bread to make a

sandwich. However, the problem is deeper than a question of arrangement of parts (and hence immune from the reservations we had over Armstrong's appeal to structured composition as an argument against mereological univocality). Since the putative monster made of ham and whatever else we like is a mereological sum its existence is guaranteed irrespective of any spatial features whatsoever, and spatial rearrangement of the monster's parts has no impact on their existence. It is just the existence – nothing else – of a monster comprising the ham and everything else which existed before or after the sandwich, which is required for the objection to bite.

The problem posed by the combined aggregative and monster objections definitely does bite, and deeply. There are some ways a classical mereologist might reply, although they are less than convincing. The first is to observe that Fine's objections can be seen to be a variation on the extravagance objection to universal composition discussed in chapter three. The problem with the monster is essentially that it is an object which must exist according to universal composition, but one which has some odd properties. As such, the same response can be levied as against extravagance: so what? All mereological sums exist, but some are more interesting than others, and the really weird ones can be effectively disregarded as being of little interest – with empirical data to support this view. This is distinctly weak, firstly because Fine's objection generalises to any composite material object whatsoever (bar the whole universe across all time, if such a thing exists), and secondly because the problem is that according to whatever temporal restriction is put in place to address the aggregative objection, the resulting monster is *part* of the material object (in the example, a sandwich). This isn't a non-standard property which can be safely ignored – parthood is what mereology is all about – and the monster is a part in the classical mereological sense, not some other sense which can be dismissed as not being genuine.

A second line of reply is that made by Sider (2007, fn. 65 in online version), which cites Fine's 'Things and their Parts' if not the actual objections made in it. There he observes that his own approach to studying parthood is distinctly 'analytic': 'highly abstract principles about parthood were formulated, and allowed to drive conclusions about particular cases' (*ibid*., p.37). This is, of course, in line with the way classical mereology is standardly developed. In contrast Fine, amongst others, adopts a more 'synthetic' approach which

'begins instead with ordinary judgments about particular cases of parthood, and erects a metaphysical edifice on this foundation' (*ibid*.). According to Sider the two approaches are fundamentally at odds, with the implication being that his own theory is immune to criticisms raised by Fine as the methodological differences between them are too great.[82] At least, this is how I read Sider's comments; if inaccurate to his intention it still makes for an interesting line of thought. But is it really plausible?

A first response on behalf of Fine: although Fine's general approach is 'synthetic' in the sense Sider describes it, the specific objections are as 'analytic' as they possibly could be. As we have seen, both are at heart *reductio* arguments which start with the highly abstract principles of mereology, apply them to specific cases, and find the results so outlandish that those principles must be revised. If this isn't a model of analytic methodology, I don't know what is. I suppose we could throw in a few washing symbols for luck?

4.4.2 Intuitions versus Evidence

I think this alone disposes of Sider's reply. It's one thing to point out that the tenor of another philosopher's approach is quite different to your own, but to conclude that as a result you can safely ignore any particular argument he makes is to engage in a bait-and-switch of the simplest variety. For those who disagree, there is a second line of response which is well worth pursuing as well. Let's imagine Sider is right that his and Fine's methodologies pass each other by like ships in the night, and that this is sufficient to make their theories and arguments incommensurable. Which are we to prefer? It should be obvious that the real situation is not a stark either/or between so-called 'analytic' and 'synthetic' methods, but we can see a clear difference in starting points. A 'synthetic' approach starts with what we have followed Poli in calling a descriptive ontology, which aims to provide a good pre-theoretical account of *prima facie* information about some domain or area of interest – in this case parthood – which the metaphysician then proceeds

[82] 'Analytic' and 'synthetic' remain in quotation marks as I am not convinced Sider's use of them here is entirely standard. 'Top-down' and 'bottom-up' respectively might be better choices but I will refrain from adding any further terminology, and stick with Sider's terms.

to systematise and refine into a formal theory. Expressing the formal theory in a formalised logic framework is an optional third step. In contrast, an 'analytic' method starts with general abstract principles about the concept(s) being studied – again, here, parthood – and once a formal theory has been developed which satisfies the metaphysician's standards it is applied to actual and hypothetical cases. But what are the 'analytic' metaphysician's standards? Logical consistency is a must (for classical mereology and its extensions discussed above, all principles must be theorems of first-order logic plus identity and one or more additional operators), but the abstract principles he or she employs have to come from somewhere in the first place. Where they come from is the individual philosopher's intuitions. However mereology developed historically (and let's not forget that set theory, on which it is based, is frequently held to have an intuitive basis), the question facing individual philosophers today using the 'analytic' approach, like Sider, is whether mereology's principles are intuitive or not. Sider takes them to be highly intuitive both individually and *en masse*, springing a comprehensive list – including some features not explicitly discussed here – almost from thin air (2007, p.20 of online version). His justification is that 'each thesis can be seen as flowing from the aphoristic conception of parts as intimate with their wholes' (*ibid*.). In other words, he finds them intuitive.

But where do 'analytic' philosophers' intuitions come from? It cannot be from specific situations or examples directly, as that would collapse the distinction with the 'synthetic' approach. They must be the result of some sort of process of selection, aggregation, mystical insight or I don't know what. I'm not at all convinced anyone else does either – one of the tremendously difficult things about relying on intuitions in philosophy, or anywhere else for that matter, is that it is nigh impossible to say where they came from, or to give them a reasonable justification. This leads on to a second difficulty: persuading anybody else that your intuitions are at all reliable or superior to their own, if they do not share them to begin with. Clashes of intuition are all too common, not least in metaphysics where the subject matter may be far removed from any shared experience or activities to help bridge the divide. I am reminded of a very short exchange at the Royal Institute of Philosophy "Being" conference (University of Leeds, 1st – 3rd September 2006) between John Hawthorne and Dean Zimmerman – both highly respected metaphysicians – concerning their contrasting intuitions. The fact it was so short is telling, as neither apparently had any

way of persuading the other. Admittedly, the clash was only over how to correctly describe sunrises, rather than some vital universal principle, but the point remains that persuading somebody that your own intuitions are correct is a tricky business, with no apparent method *in general* available to do so. Thirdly, however communicable intuitions may or may not be, philosophers' ones are – to borrow a phrase (used in conversation) from Paul Griffiths – notoriously corrupt. Having spent many long afternoons, months or even decades mulling over some thorny issue, no-one could reasonably expect any intuitions a philosopher might have had about the matter to remain unchanged across this process. Using intuitions as a primary basis for philosophical theorising is a risky business: they are obscure, difficult to justify, and liable to be infected with the very theories they are supposed to inform. Admittedly, everyone has to start somewhere, but there is a vital difference between using intuition as a starting point and using it as evidence to support one's conclusions.

As a response to Sider, this is essentially an exhortation to jump ship and join the 'synthetic' crew. Two responses on behalf of 'analytic' philosophy spring to mind: first, that it isn't primarily based upon intuitions plus logical principles, and second that its foundations are no less shaky than the 'synthetic' alternative. The first might just be true, but the use and abuse of intuitions in philosophy, including metaphysics, is very widely recognised – and it is quite transparently in play in many approaches such as Sider's (2007). I also know of no serious alternative explanation of how general philosophical principles are derived without any specific evidence. In the absence of any convincing alternative, the first response falls far short of being convincing. The second response has more force, as a general method of deriving philosophical principles from specific cases is not entirely clear or straightforward either. In the following section I will illustrate how it can be done through presenting an outline of Fine's own theory of composition – of rigid and variable embodiment – and how it might be supported using empirical evidence against criticisms raised by Koslicki (2007). In contrast to the 'analytic' approach, the 'synthetic' approach is superior because it is – or at least can be – based primarily on testable evidence from which theories are derived, rather than on abstruse theories into which testable evidence is shoehorned.

4.4.3 Rigid and Variable Embodiments

Fine's theory of composition is, as he readily admits, incomplete as a general theory of parthood. Nevertheless, it is sufficiently well developed to give a good indication of what the finished product would be like. With regards to material objects, Fine suggests two parthood relations (leaving open the possibility of other relations in other domains), one tenseless and one tensed. He calls them rigid and variable embodiment respectively. In this section we will outline these two principles of embodiment, apply them to our three problem cases, and assess and respond to recent criticisms raised by Koslicki (2007). We will conclude that Fine's theory is in no worse position to address the problem cases than a supplemented classical mereology. Given our emphasis on the value of empirical data in ontology, this means Fine's theory is preferable.

Rigid embodiments exist if and only if some objects *a, b, c* exist and a suitable relation *R* holds between them; the embodiment itself is taken to be a primitive relation (just as proper parthood, or some alternative, is in classical mereology). Following Fine, we can express this as <a, b, c, … /R>. The objects *a, b, c* and the relation *R* are all timeless parts of the rigid embodiment. This produces some significant differences between rigid embodiments and mereological sums. Unlike mereological sums, rigid embodiments only exist when all their parts exist (like sums, they also exist wherever their parts exist), and they are distinct objects over and above their parts – there is absolutely no question of rigid embodiments being identical to their parts, either literally or by close analogy. Like mereological sums, rigid embodiments admit no change in their parts.

To return to the ham sandwich, a rigid embodiment exists whenever the two slices of bread and the ham all exist, and when a relation *R* holds between them of being suitably arranged with the ham between the bread. This neatly sidesteps the aggregative and monster objections as it is a conjunctive parthood relation. Admittedly, as a full analysis of what it takes to be a sandwich this is too schematic – surely the sandwich can accept some variation in parts – but it serves to illustrate the motivation behind rigid embodiment, as well as how the principle works. Given the sheer quantity of material objects and relations which hold

between them, the number of rigid embodiments in the universe will be vast. It is not clear to me whether all properties and relations are suitable to produce principles of rigid embodiment or not. Fine imposes no restrictions, and while for relations he only mentions paradigm cases such as connection and spatial arrangement, properties which produce rigid embodiments include descriptions such as being an airline passenger, mayor or judge, an action or even a trope. If classical mereology is exuberant, Fine's theory is breathtaking in its sheer proliferation both of parts and wholes, and of *sui generis* parthood relations. Embodiments according to different principles may also share exactly the same spaces and times – some group of objects may share multiple relations, each of which is suitable to produce a rigid embodiment. After all, /R is a different relation to /P or /Q, assuming the relations R, P and Q themselves are not the same, and the principles of embodiments are included amongst their parts.

Rigid embodiments are, however, clearly inadequate to deal with the composition of material objects in general – mereological essentialism notwithstanding, one of the most remarkable features of many material objects in the current context is that they can and do change parts. Take a river, for example, whose water constantly flows along its path. This is where variable embodiments come into play; they are made up of a series of 'manifestations' at different times as well as a 'suitable function or principle' (Fine 1999, p.69) of embodiment which goes from times to objects. Unlike rigid embodiments, from these manifestations variable embodiments inherit properties such as colour, size, shape and so on. The manifestations of a variable embodiment are temporary parts, and may be rigid embodiments, or they may be further variable embodiments. So the water which flows through the river is analysed by Fine as a variable embodiment: at each moment the river flows there is a rigid embodiment consisting of different specific volumes of water for each time, united by some principle or function which picks out just those rigid embodiments (the variable embodiment's manifestations) at the right times. Just what the principle or function would be is not entirely obvious, but assuming there is only one river there would be a unique principle of (variable) embodiment which produces it, and it would presumably incorporate location and material composition. A body of water shouldn't count as a manifestation of a river if it is in completely the wrong place, and a volume or mercury or lard wouldn't count even if in the right place; neither are the right kind of stuff, and the

latter don't even flow, at least not at British Summer temperatures. While some impurities must be allowed, if it isn't (mostly) water, and if it isn't in a river bed, it isn't part of a river. Except that the bed might be counted as part of the river, and Fine's theory counts principles of embodiment as parts too. It would be nice to be able to make a short, pithy statement about the composition of material objects, but matters are simply too complicated. Slogans are for sloganeering, not for real work.

In essence, this gives us the core of Fine's theory of rigid and variable embodiments. He does present the theory in slightly more technical terms than I have reproduced it here, but nothing of any use to us is lost through an informal presentation. The only real value for use of washing symbols with classical mereology and its extensions is that everyone else is at it, so joining in is useful for preserving continuity more easily with others' work. Fine's theory is as-yet incomplete and only partly formalised, so there is no need to do the same, at least not yet. Before examining some criticisms of the theory we should get a better idea of how it works by applying it to our three problem cases. Since Fine explicitly discusses the example of a car, let's start with number two: in what sense, if any, has a car changed if we place its engine on the back seat? Cars are variable embodiments whose principles link a series of manifestations across the period of its existence. However,

> the current manifestation f_t of the car is a temporary part of the car (at *t*). But the manifestation f_t is itself a rigid embodiment a, b, c, ... /R, namely, the engine, chassis, and body in a certain automotive relation ... the engine, chassis and body are timeless parts of the rigid embodiment *ft* and hence ... are temporary parts of the car ... But the engine itself is a variable embodiment and so, in a similar manner, it will have temporary parts, which will be temporary parts of the car.
>
> Fine 1999, p.70

Cars are pretty complex beasts, and not just mechanically. One material object is made up of a sophisticated hierarchy of temporary parts, some of which have their own parts temporarily, and some timelessly (incidentally, if *f* has a temporary part *x* which has a timeless part *y*, then *y* is a temporary part of *f*. This much should be quite intuitive). Earlier we discussed what impact moving the engine to the back seat would have on the car itself, and canvassed three options: the car would still exist, albeit in a non-functioning state, the car would cease to exist and a new one would be created if the engine

were replaced where it should be, and the car would cease to exist but come back into existence were the engine to be replaced. These three options correspond to variations in the 'automotive relation' which acts as the principle of variable embodiment for our car, so as with mereotopology, Fine's theory of embodiments does not legislate which option is correct but leaves room for all of them. Again, this is a virtue. It seems pretty clear that the theory of embodiments deals equally well with our second problem case as mereotopology does, with the same number of primitive relations and a less ornate conceptual framework.

Fine's theory applies equally well to our slightly simpler first problem case, explaining the difference between a cairn and a pile of stones. A cairn, on Fine's view, is a variable embodiment made up of a series of manifestations. Each manifestation is itself a rigid embodiment made up of a particular aggregation of stones at a particular time. While a specific pile of stones at time *t* may not vary in its parts, the cairn which is embodied in that pile at time *t* may change in its parts over time. This accords with common usage of the word 'cairn' – by definition they gain additional stones as parts over time, and the removal of a single stone would not be sufficient to destroy one.

A similar answer again can be given to our third problem case, that of what sense half of an uncut apple is part of the apple. An apple is a variable embodiment composed of a series of manifestations which are themselves rigid embodiments. Each of these manifestations would be a whole object at a specific time; the uncut half of the apple is a timeless part of the manifestation which has 50% of the volume or mass (or both) of the manifestation. As such, the uncut half of the apple is a temporary part of the apple, but exactly which part will depend *inter alia* on when we are considering the apple.[83] This is intuitively correct: apples grow, so half an apple at time *t* will not necessarily be the same as half at t^{+1}.

So Fine's theory of embodiment is capable of dealing effectively with all the problem cases we have considered, and as such it can be considered at least as acceptable as classical mereology when supplemented in the ways described above. Our discussion of Fine's theory has been relatively brief in comparison to that of mereology; this is a reflection of

[83] Other relevant factors include how we specify that half, e.g. left or right relative to a point of reference, internal or external and so on.

the current literature and the relative novelty of Fine's view. Whether it will in the future be discussed in as much detail as classical mereology has so far remains to be seen. Either way, our endorsement of the adequacy of Fine's theory only holds provided the theory suffers from no serious objections. One such objection is raised by Koslicki (2007), who argues against Fine's theory on the grounds that it creates a superabundance of objects, and ones which have a questionable ontological status.

4.4.4 Superabundance of Objects

Fine's theory clearly requires that there be a very large number of objects in existence at any time, and that very many of them will share their parts entirely with others. Consider a motor car – we have seen that it is a variable embodiment composed of rigid embodiments, some of which are themselves composed of variable embodiments. It is reasonable to suppose that a complete description of the composition of a car, or any material object, will involve very many levels of embodiment, all the way down to fundamental particles (or whatever the 'ground' level of embodiment is – assuming that there is one, of course). Considered as a hierarchy with the material object at the top, any one object will share all its parts with every object at a higher level than it. Considered in isolation, this makes many of these parts rather curious objects. They are not whole material objects themselves (though some will be, depending upon what object is at the top of our hierarchy), and in keeping with the potential parts doctrine their very existence and identity is dependent upon the whole they are part of. To some ways of thinking this makes them very curious objects indeed – a far cry from the sorts of things we are directly acquainted with through experience.

The second thread of Koslicki's objection is the sheer number of objects which, on Fine's view, compose a single material object. A typical car is made up of thousands of discrete components, but dwarfs in complexity compared to living organisms. A human body of 70kg is made of something in the vicinity of 10^{25} atoms (a figure widely attributed to Michio Kaku, though I have been unable to verify this) – that's 100,000,000,000,000,000,000,000. When

we add all the molecules which take these atoms as parts, then cells, organelles, neurons, synapses, organs, capillaries, blood vessels, etc the number of objects which go to make a human being will on Fine's view be absolutely enormous. And that is a single human being. Koslicki's real objection is the combination of this abundance with the ontological queerness of the majority of the objects.

However, Koslicki's concerns are ill-founded, for exactly the same reason that concerns over the exuberance and extravagance of unrestricted mereological composition are. There is a wealth of empirical data concerning both adults and infants which indicates that we show selective interest to certain types of properties in our environment, to the exclusion of others. Starting in very early infancy we show clear sensitivity to quantitative properties of objects such as number and relative size – but not to qualitative properties such as absolute size, shape or colour – and build upon these sensitivities through adult life. This provides strong support for a 'detectivist' view in ontology according to which there are a great many objects and properties in existence, those we are most directly aware of being detected from this mass in accordance with the functioning of our sense and our own interests. We will return to the topic of infant development in chapter five when we consider Spelke's core knowledge hypothesis in connection with commonsense psychology; for now we can see that given the empirical support for a detectivist view in ontology, Koslicki's objection to Fine does not bite.

4.5 Parts and Wholes: Mereology or Embodiment?

We have seen that classical mereology alone provides inadequate explanations of all of our three problem cases. We have also seen that a suitably supplemented mereology can plausibly explain each of the cases. The difference between a cairn and a pile of stones can be explained by incorporating a topological concept of connection plus a causal stability requirement. The sense in which a car has changed if its engine is placed on its rear seat can be adequately explained by supplementing mereology with topological connection plus a functional requirement to capture the concept of working parts. The distinction between a

whole apple and its uncut half can be captured by supplementing mereology with the topological concept of strong self-connection (and maximally strong self-connection). This gives us good reason to suppose that classical mereology, with considerable additional resources added, can provide an adequate account of parthood and wholeness. However, we have also seen that certain assumptions need to be made in order to agree with this result. In particular, mereological extensionality lacks motivation and is difficult if not impossible to accept as a general principle unless one is already a perdurantist. Mereological reflexivity, antisymmetry, transitivity, extensionality and unrestricted composition are all defensible. Mereological uniqueness is shown by Fine's objections to be false.

By contrast we have seen that despite its relative newness and limited coverage in the literature so far, Fine's theory of rigid and variable embodiments can provide an adequate explanation of each of our three problem cases. All material objects are hierarchies of variable and rigid embodiments, made up of both temporary and timeless parts. The difference between a cairn and a pile of stones lies in a cairn being a rigid embodiment whose time-linked manifestations are themselves piles of stones, which are in turn rigid embodiments composed of individual stones. A car is a variable embodiment composed of manifestations of components at specific times, with those components being either rigid or variable embodiments depending upon their function. Moving a car's engine disrupts the rigid embodiment which is a manifestation of the car at that time, and hence Fine's theory can accommodate a range of conclusions about the impact that will have on the car itself. An uncut half of an apple is a rigid embodiment which is itself a proportion of the rigid embodiment which is the manifestation of the variably embodied apple at a particular time. This gives us good reason to suppose that Fine's theory of variable and rigid embodiments can provide an adequate account of parthood and wholeness, without the need for supplementation on any large scale. However, as with mereology certain assumptions need to be made in order to accept Fine's theory. In particular, since two objects may share parts without being identical, it entails that mereological extensionality is false. This is germane to endurantists, and unacceptable to perdurantists.

We have also witnessed the failure of Casati and Varzi's attempt to incorporate into an extensional framework the intuitions which underlie theories such as Fine's, through their countermereotopology. Had they been successful, this would plausibly have undermined Fine's theory despite its successes, but that has been shown not to be the case. We are left with a choice in how to explain the part-whole relation: supplemented mereology, or the theory of embodiments? Either is acceptable, depending upon one's prior theoretical commitments. The choice is comparable to asking a vegetarian if they would like beef wellington or boiled tofu; the beef may be excellently prepared, but their decision has already been made (even if they had never actually considered it until now). Because of their commitment to a particular view of the ultimate nature of the universe, perdurantists will choose mereology. Because of their prior ontological commitment to a different view of the ultimate nature of the universe, endurantists will choose Fine's theory (or a comparable alternative).

I take it that this is a vindication of (1) the distinction between descriptive and formal ontologies, and of (2) the method Sider calls synthetic ontology. Every substantive decision in one's formal ontology turns upon the descriptive ontology which it is an analysis of. That is not to say the decisions *only* turn on the descriptive ontology, but they are always influenced by it. This has been shown most clearly in the confused debate over mereological extensionality. Attempts to provide empirical counterexamples to extensionality have always been doomed to failure, because whether the formal principle of extensionality is acceptable or not follows from one's descriptive ontology (in particular, whether it is 3D or 4D). This descriptive ontology concerns the metaphysical status of all material objects whatsoever, so hunting for an empirical example to settle the debate over extensionality is futile.

However, empirical data is absolutely central to forming a sound descriptive ontology in the first place. The ultimate goal of formal ontology is a rigorous and systematic understanding of the general concepts which are central to our actual thoughts and practices: concepts like parthood, identity, dependence and representation. A descriptive ontology which does not present an accurate account of our actual thoughts and actions will necessarily produce a distorted and misleading formal theory. My own view is that in the case of parthood, the

perdurance view combined with treating material objects as 'individuals' has done exactly that. Mereology is an abstruse and contorted system which can only be rendered functional by considerable gerrymandering, and even then only functional by the standards of those who are already inclined to accept it. But we have not argued for that particular view here: nothing said has provided evidence that perdurance is mistaken, and with sufficient gerrymandering mereology has been seen to work. The point being made is methodological: without a sound descriptive ontology, any attempt at generating a formal analysis will produce a mistaken theory.

4.6 Concerns and Methods of Ontology, Yet Again

The first concern of ontology, on the synthetic view we have advocated and used, is to produce a sound descriptive ontology, a full and accurate characterisation of the phenomenon under consideration. The second concern, which follows only once the first is addressed, is how to analyse and structure this characterisation to produce a rigorous and systematic formal theory. In considering classical mereology and its variants, and contrasting them with Fine's theory, we have thoroughly demonstrated the validity of this approach as well as the problems which can arise if it is not followed.

The methods of synthetic ontology have also been effectively demonstrated through our analysis of mereology and its variants. The arguments we have considered are highly *a prioristic*, and it is often considered to be one of the defining features of ontology that it should be so. For instance Guarino writes 'formal ontology can be intended as the theory of *a priori distinctions* ... among the entities of the world ... [and] ... among the meta-level categories used to model the world' (1995, p.5 of online version). Husserl's conception of formal ontology, from which the contemporary use draws inspiration, considered it to be based upon analytic *a priori* truths, with his material ontology (roughly analogous to our descriptive ontology) based upon the synthetic *a priori* (for further details see e.g. Poli 1993, Smith 1989). It would be a mistake to conclude that for this reason ontology is divorced from empirical concerns. Time and again we have seen that empirical data – whether drawn

from science as in our use of optical isomerism in relation to extensionality, or personal experience as in the discussion of door handles in relation to transitivity – plays a vital role in informing the *a prioristic* arguments which make up the heart of ontological study. The central morals to be taken from our excursion through ontology, then, are quite simple, though not uncontroversial:

1. Formulate a sound descriptive ontology which accurately characterises the phenomena being considered, prior to developing a rigorous philosophical framework.
2. Both the descriptive ontology characterising a phenomenon, and the formal ontology which analyses it, should be guided and informed by empirical data (and not intuition or personal preference).

It is, in my view, only fair to comment that both of these morals have been utterly disregarded by almost all of the literature on classical mereology. Advocates of mereology will most likely reply that their descriptive ontology is correct and (2) is false. My argument against this lies in the effectiveness of my approach in exploring mereology and the part-whole relation. These morals are justified to the extent that I have succeeded in clearly and accurately analysing the relevant issues and how they can be addressed; I take it that I have done so, but leave it up to the reader to draw their own conclusion. In the chapter five we will apply these morals to the Representational Theory of Mind, and argue that by doing so we can discover good reasons to conclude that mental representations do not exist. First, the descriptive bases of RTM is fundamentally flawed. Second, by far the most attractive aspect of RTM – causal semantics – can be readily preserved even if mental representations do not exist. Together these facts present compelling evidence that there are no such things as mental representations.

4.7 Metaphysics and Mental Representation, Again

We have already seen that RTM is an ontological theory, in the sense that it stipulates the existence of mental representations. If mental representations, with their characteristic properties and relations, did not exist then all varieties of RTM would be false. Hence, RTM is ontologically committed to their existence.

RTM is also a formal theory, in precisely the sense articulated by Poli. Its goal is to analyse the phenomenon of intentional thought to produce a rigorous and systematic theoretical framework which explains its key features. This is precisely the same as mereology, which analyses the phenomenon of parthood to produce a rigorous and systematic theoretical framework which explains its key features. This close analogy is easily missed as mereology is standardly presented in a logical format and RTM is not, but the point nevertheless holds.

We have seen that any formal theory must be developed on the basis of a descriptive ontology, a characterisation of the phenomenon being studied. For classical mereology the descriptive ontology is one of treating all material objects as discrete individuals (in Leonard & Goodman's sense) in a four-dimensional framework. The descriptive ontology which RTM analyses is commonsense psychology. Recall that Fodor states that for a mental representation to have the meaning it does just is for a propositional attitude to have the content it does, and Millikan requires every intentional thought to have a corresponding propositional attitude, as well as a corresponding representation. Fodor directly endorses the view that the majority of everyday cognition consists in commonsense psychology, and while Millikan does not discuss commonsense psychology, her commitment to propositional attitudes and their vertical relations with mental representations requires she adopt a comparable view, if not precisely the same. Commonsense psychology is a descriptive characterisation of intentional thought which tells us that it consists in propositional attitudes such as beliefs and desires, and the uses to which they are put, specifically explanation and prediction of behaviour. RTM analyses that description by postulating the existence of a token mental representation for every propositional attitude, with the horizontal relations between propositional attitudes being explained by the horizontal

relations between token mental representations. There can be no reasonable doubt that RTM is a formal theory in Poli's sense, and that its descriptive ontology is commonsense psychology.

By applying our two morals to RTM, we can see that the primary concern in assessing its validity as an ontological theory is whether its descriptive basis is sound. Only once this has been established can we then tackle the second question of whether its analysis is sound. In both of these we should take great care that our arguments be led by empirical data, as well as sound reasoning, and avoid as far as possible personal sentiments or intuitions.

It might be hoped that our study of parthood could be applied directly to mental representation, and thereby deliver either a refutation of RTM or an endorsement of the way it characterises composition. Unfortunately this is not possible, because the stipulations made by RTMs about how representations are instantiated are simply too vague and imprecise. They amount to two claims: mental representations are physical states or processes in the brain, and they are internally structured. While in recent years Fodor has begun to explore the ways in which mental representations may be structured (see e.g. his 2003 pp.37-8 & 2008 ch.6) this has been limited to differentiating between conceptual and non-conceptual content. This sheds no significant light upon how representations themselves are structured in general, not least because – as we shall see in chapter six – his method of doing so fails. While on the face of it Fodor's CTM might seem to be a substantive claim about the nature of compositionality – that horizontal relations can be accurately modelled using a Turing machine – it tells us almost nothing of any substance about the structuring of representations themselves. This is because any process can be modelled to any arbitrary degree of accuracy by some computational process; even if CTM is correct, it tells us nothing about how representations are *actually* instantiated, and hence nothing more about their structure than RTM does. To put it another way, any given computational process is multiply realisable to a massive degree, so the process itself tells us nothing about the architecture it is being realised by.[84]

[84] For example, the Folding@Home project (http://folding.stanford.edu/English/Main) at Stanford University computes the ways in which protein molecules fold by exploiting unused capacity in domestic computers. There are currently 350,000 computers contributing to this 'distributed computing' network, but the same

The claim that mental representations are internally structured arguably presupposes the falsehood of classical mereology; while we have seen that mereology can explain apparent counterexamples to extensionality by appealing to non-mereological properties of the objects, it is doubtful whether that strategy could be applied here. The reason is that unlike molecules, for example, mental representations are not a class of entities which have been empirically discovered complete with a wide range of features. Rather, their existence and nature are stipulated by RTM which simply claims they are internally structured physical processes or states which participate in the horizontal relations described in chapter one. There is no plausible candidate for a non-mereology property for representations which could explain why their apparent internal structure is illusory, and adding one would constitute a significant development of the theory. Even if this were done, compositionality as understood by Fodor and Millikan violates extensionality. Unlike Fodor's view of associationism, which fails to differentiate between 'cow brown' and 'brown cow', RTM distinguishes ways parts are arranged. RTM and mereology cannot both be true.

Fine's theory of rigid and variable embodiments, however, is compatible with the internal structure mental representations require, but the imprecision of RTM's claims precludes any direct analysis using Fine's theory. Are token representations rigid or variable embodiments? Are their manifestations themselves composed of rigid or variable embodiments, or a combination of both? Should representations properly be considered as objects, or else as manifestations? There is simply no answer for any of these questions to be drawn from RTM itself, and so any attempt to answer them would be pure speculation. If we had time and space available here we could perhaps suggest some answers, but doing so would be to put words into the mouths of the advocates of RTM, not analyse the claims that they make. Regrettably, in the current state of development of RTM there can be no direct application of any metaphysical theory of parthood to mental representations. In the future such an analysis may be possible, and if so would certainly be worthwhile.

---

computations could be performed on a ticker tape machine. It would take something in the order of trillions of years longer, but in principle it would work.

This in no way diminishes the relevance of our study of mereology and non-mereological composition in relation to mental representation, or undermines the conclusions we will draw. Our discussion of parthood paradigmatically illustrates appropriate (and inappropriate) method in metaphysics in general and formal ontology in particular. As a formal ontological theory, based upon the formal concept of representation, RTM needs to be compatible with, and preferably make use of, the sound methodological principles identified. As we shall see, the theory fails spectacularly in this regard and so loses any plausible claim to credibility or truth.

In the next chapter we will question whether commonsense psychology is a sound descriptive characterisation of everyday intentional thought, and we will see that there are strong empirically-supported arguments to believe that it is not. Because of this RTM is akin to a house built on mud; whatever attractions the theory may present, it is based upon inadequate foundations and so necessarily fails as an analysis of intentional thought. Analysis of a false characterisation of a phenomenon – however sophisticated the analysis may be – is doomed to failure. This provides a strong argument that mental representations do not exist.

The root of the problem with RTM is that it cavalierly imposes a theoretical framework upon the phenomenon it seeks to explain, rather than examine the phenomenon of intentional thought carefully then develop a suitable analysis based upon the results of this examination. While some aspects of RTM are attractive – in particular, its causal semantics – it is methodologically unsound. In chapter six we will argue that, by making some minor and naturalistically acceptable adjustments to their descriptive ontologies, both Millikan and Fodor's semantics can be seen to function perfectly without any need or role for mental representations. Mental representations simply do not exist.

5. Commonsense Psychology

So what sort of descriptive ontology is presupposed by representational theories of mind? How do they characterise our ability to think, prior to formalising it in terms of relations between representations? As we saw in chapter one, advocates of RTMs almost universally take thought to consist primarily of propositional attitudes which relate thinkers to mental contents, and uses to which these propositional attitudes are put. This might sound a little strange, but it is utterly endemic in contemporary philosophy of mind (so, then again, might not sound very strange at all to anyone reading this). In particular, the paradigm cases of thought are held to be predicting and explaining what will happen around us and doing so in terms of beliefs and desires, as well as other propositional attitudes. This process of prediction and explanation in terms of beliefs and desires is referred to as commonsense (also 'folk', 'everyday' or 'naïve') psychology, and is supposedly what most of us are up to most of the time. That is the descriptive ontology presupposed by representational theories of mind. It is widely regarded as obviously true, in some quarters as beyond question. In this section I will argue that it is not obviously true, but should be regarded as a highly questionable theory which is most likely false. The arguments typically used in favour of commonsense psychology as a description or analysis of how people think are unconvincing, and there is a wealth of empirical evidence which directly undermines the view.
Before we go any further, a few disclaimers:

One, aside from my evaluative remarks at the end, the above description of 'commonsense psychology' is quite uncontroversial and is universally accepted in the literature (for more details refer back to chapter one). The only unusual aspect of my presentation is applying the term 'descriptive ontology', but given our earlier discussions of the meaning of the phrase, and that RTMs should be viewed as formal ontology, I think this requires no further support.

Two, I don't think anyone serious has claimed that *all* thought consists in propositional attitudes (anyone who believes in non-propositional mental content – and many do – must *ipso facto* disagree with that idea). The claim is rather that *to a large extent* thought consists

in fomenting propositional attitudes, and that they constitute the core of our mental faculties. Fodor in particular is absolutely explicit on this point in many, many places, and as we have seen Dretske frames his discussion in these terms. Millikan explicitly acknowledges their centrality to thought, for example when she says 'intentional representations always come with propositional attitudes attached' (2004, ch. 6; page 6 of online version).

Three, nobody seriously believes that the only propositional attitudes people use are belief and desire. Again, these are the paradigm cases. They are discussed almost to the exclusion of all others, in particular belief, but it is widely accepted that people do hope, fear, intend, and do a wide variety of other things as well. Some have questioned whether we primarily or exclusively use propositional attitudes to predict and explain (e.g. Morton 1996), but once more these are more-or-less universally treated as paradigm cases.

These disclaimers aside, advocates of representational theories of mind take our thought to consist primarily in what they call commonsense psychology, in predicting and explaining behaviour in terms of beliefs and desires. An infamous example of this type of reasoning is:

> *X* believed that it was raining and desired to stay dry, so he took an umbrella when he went outside.

Changing the tenses of the verbs readily converts this explanation into a prediction. The question is, how good is this as a description of the ways in which people actually think? It is widely regarded as being obviously true, so much so that it is often seen as beyond reasonable doubt. Botterill and Carruthers, for example, opine that 'we cannot help but think of each other in such terms' (1999, p.10). Matthews even goes as far as to characterise this view as being so entrenched that it 'has for its proponents the status of something approaching an a priori necessary truth' (2007, p.7). Certainly for decades it was rarely questioned, but in the last ten years or so it has increasingly come under criticism.

Fodor once claimed that 'the predictive adequacy of commonsense psychology is beyond rational dispute' (1987, p.6), i.e. it is an unquestionably accurate account of how people think. I will argue in this section that:

1. The adequacy of commonsense psychology as an account of how people think is not beyond rational dispute, and
2. It is almost certainly an inadequate account.

But before we turn to the arguments, there are several potential confusions in the literature which need to be addressed first to make the later discussion as clear as possible. These concern defining the meaning of the term 'commonsense', particularly in terms of its relationship with other non-commonsense activities such as scientific or other theoretical reasoning.

5.1 What is 'Commonsense'?

A simple point of inadequacy in the 'commonsense psychology' theory is this: not all thinking is about people, thoughts or behaviour. It seems obvious that someone looking to describe how thinking works should settle on broader cases than just thinking about people. There have been attempts to articulate commonsense or 'naïve' physics, for example (see the papers collected in Hobbs & Moore, and discussion in Smith & Casati). It seems odd to focus on one domain of thought, but perhaps it is a forgivable shortcoming. After all, philosophers of mind spend much of their time thinking about thinking, so why not imagine other people do the same? We will return to this point at the end of the chapter. To be fair, a case might also be made for the majority of our thought and interaction to take place in a social environment, one where other actors and their thoughts take centre stage. To that extent, commonsense psychology might not be the whole of a description of how people think, but might well be a central part, and hence a good place to start.

Unfortunately, there are a number of points of potential confusion regarding the use of the word 'commonsense'. It is often claimed that there is a commonsense or everyday world view, which scientific and other theories depart from to varying degrees, and which is adverted to by talk of 'Joe Average', 'A. N. Other' or just occasionally 'the man on the Clapham Omnibus'. But under closer scrutiny, it is horribly difficult to say exactly what this

view of the world actually *is* (a point not lost on early phenomenologists). It's hard enough to even specify what the term 'commonsense' in this context actually means. An easy enough contrast is with scientific psychology – or scientific physics, physiology or whatever else – in virtue of the two apparently making contrasting claims about similar or the same subject areas. But the relationship is far less straightforward. Consider these five cases:

1. Many or most people would agree that some objects are dense with smooth surfaces. Modern physics tells us that they are not just pitted, but almost entirely composed of empty space.

2. People regularly make choices and predictions which are 'sub-optimal' in comparison to probabilistic models of the situation and possible outcomes.[85] There are numerous well-documented examples, including the base rate fallacy and the prisoner's dilemma (see e.g. Bar-Hillel and Kuhn respectively).

3. Leibniz and Newton disagreed over whether space is nothing over and above relations between objects, or if it is more like a container in which objects are located. If you simply want to walk from *A* to *B*, the difference is irrelevant.

4. A boundary can be defined by the path of a river. But as the river meanders, does the boundary also move? As we have seen in chapter four, attempting an answer to this simple question rapidly leads to complex and highly abstract theoretical distinctions between boundary types and conceptions of space.

5. Technical theoretical terms can be absorbed into common usage over time, with a closely similar meaning, e.g. 'gene'.[86]

1 and 2 tell us that scientific or other theoretical claims may correct or supersede commonsense judgments in some way: 'the surface may feel smooth, but it's *actually not*…' This suggests that commonsense judgments and/or reasoning may be flawed and inaccurate. This view is promoted in particular by Paul Churchland (1981, 1988), who is a

[85] It is reasonable to question in what sense people are 'wrong' in making probability judgments which disagree with mathematical models; perhaps they are employing a generally successful mode of reasoning which comes unstuck only in unusually abstract contexts. Unfortunately we don't have room to pursue the idea here, although I suspect it probably is close to the truth.

[86] 'Gene' is quite a nice example, as in more-or-less all contexts it is used to mean a hereditary unit. More technically it is often defined as the portion of DNA which codes for a single polypeptide (with chains of polypeptides forming proteins or RNA). In everyday usage this level of detail is usually irrelevant, and often unknown.

realist about commonsense psychology to the extent that he believes people do in fact employ it, but an eliminativist in that he argues it should be replaced *tout court* by a more scientifically accurate alternative. I don't intend to discuss this view in any detail, though some of the discussion below is relevant, since as we will see there are good reasons to doubt that people do employ commonsense psychology in the first place. Eliminating something which doesn't exist is, frankly, a waste of time.[87]

3 reminds us that commonsense considerations are linked to particular concerns or domains of interest, and this may limit the extent to which they are susceptible to being superseded by scientific or other non-commonsensical views. There is an asinine witticism which I have heard various versions of, to the effect that "according to science bumblebees/hummingbirds can't fly", and which – I imagine – is designed to show the limitations or foibles of scientists. Usually it only reveals something about whoever utters it, but we can adapt it to make a point here about the 'world of commonsense'. Personally, I don't have much sympathy with the view that there is a single unified commonsense view of reality, but there are widespread ways of looking at things based upon our day-to-day concerns. These ways of looking at things are highly resistant to revision, even in the light of well evidenced or argued reasoning. If a scientific theory were to tell us bees cannot fly, so much the worse for the theory because it is evidently false. We all know bees can fly; we've seen them at it enough times. Science does tell us that walls are not solid (at least in an absolute sense), but few people who believe this try to walk through them. Were an eliminative project such as Churchland's to be realised, it is highly doubtful whether it would achieve mainstream use. This point is hardly news; in a similar vein, Hume's academic scepticism led him to deny that we could arrive at the idea of cause and effect by observing

[87] I can't resist a quick and dirty argument about why commonsense and science are *not* straightforwardly in competition. The fact is, if 'commonsense' reasoning exists at all, many people are engaged in both. Even someone who is convinced that everything which happens has a complete physical explanation, one which exhaustively determines future events and actions, is still likely to ask his wife for a cup of tea rather than just wait and see if one arrives – even if, when questioned, he admits that on scientific or philosophical grounds he doesn't believe he is capable of altering what will happen anyway. Determinism does not entail quietism, and the fine ideas which some theorists propound 'in the lab' (or, for philosophers, 'in the pub') don't necessarily filter down to every aspect of their lives and conduct (I suppose some might take the Kantian line that determinism only holds in the phenomenal realm, but many scientists or other theorists will not be so familiar with the 'calamitous spinner'). Without wishing to be unfairly *ad hominem*, this disparity seems particularly true of some philosophers who make remarkable claims about the ways in which 'commonsense psychology' works or doesn't work, and then act in a way completely at odds with their own description.

and reasoning about ‘constant conjunctions’, but he always left by the door rather than the window.

1, 2 + 3 strongly suggest that commonsense and scientific or theoretical judgments overlap to some extent, sharing at least some of the same subject matter, but are unlikely to be in direct competition in all cases, and perhaps less likely to be supplanted entirely by the other. Although I have used non-psychological examples for simplicity, there is absolutely no reason to doubt that this includes psychological concepts as well.

4 + 5 further complicate the relationship between commonsense and scientific/theoretical reasoning. As in the case of meandering boundaries, simple ‘commonsense’ concepts when examined can rapidly lead to highly theoretical ones, and there is no clear boundary between the two – at least not one which is likely to be specified in ‘commonsense’ terms! But there can be no doubt that boundaries are a feature of day-to-day life, so they certainly aren’t just recognized by highly theoretical enterprises (what could be more practical than drawing a line in the sand?). How and when one leads to the other is difficult to give a principled answer to, and is further complicated by the fact that scientific and other technical concepts such as ‘gene’ do cross over into common usage. There are plenty of psychological examples of this phenomenon – catharsis, denial, repression, consciousness, projection, schadenfreude, bad faith, authenticity (in contemporary vernacular under the guise of being ‘real’), ego and id all spring to mind, to name but a few. Commonsense psychology – if it exists at all – develops and adapts in its content over time, even if, as often claimed, it is ‘stagnant’ in its fundamental structure (see e.g. Churchland 1981 p.74, Clark 1987 p.146, Stich 1983 p.229). To what extent ‘commonsense’ concepts might pass into scientific or other theoretical usage I’m not sure; although several terms coined by science fiction writers have entered technical use, including robotics, (computer) viruses and worms. Arthur C. Clarke invented geostationary communication satellites in a letter to a magazine (1945), but I’m doubtful whether this really counts as commonsense influencing science.

This is hardly an exhaustive discussion, but just a few examples highlight the need for considerable care in contrasting commonsense with science or other technical or abstract enterprises. Are scientific and commonsense psychology in competition; do they address the same subject matter; might one be replaced entirely or in part by the other; how do the two interact? Apart from the last question these are fairly well recognized in the literature, primarily through debates over whether commonsense psychology is implemented in the form of a theoretical structure of some sort (the 'theory theory') , or by using ourselves as a model by which to understand others (the 'simulation theory').[88] The debate between the two, and the beginnings of a consensus in favour of a hybrid model, is nicely captured in the collections edited by Carruthers & Smith, and Davies & Stone (1995a & 1995b).

As we have seen, comparing commonsense with scientific or theoretical disciplines raises a number of thorny questions. As far as I am aware there is no definitive set of answers in the literature as of yet, or indeed a reliable consensus worth taking seriously, but at least the issues the questions raise are generally recognized and addressed explicitly. In a moment we will consider another which, sadly, is not. But there are other ways to characterise commonsense than in contrast with more theoretical or technical activities. Gibson makes reference to size relative to people and animals, observing that:

> [t]he world can be analyzed at many levels, from atomic through terrestrial to cosmic. There is physical structure on the scale of millimicrons at one extreme and on the scale of light years at another. But surely the appropriate scale for animals is the intermediate one of millimeters to kilometers, and it is appropriate because the world and the animal are then comparable.
>
> *1966, p. 21*

Extending this view to all five senses, we could suggest that the subject matter of commonsense psychology, as well as other putative commonsense disciplines (is 'commonsense discipline' an oxymoron?), is what can be seen by the unaided eye, felt by touch, readily heard and so on. This too is far from perfect – in glancing through binoculars, has someone left the world of commonsense behind? It seems unlikely. For a powerful

[88] Actually, this characterization of simulation theory is a bit crude – it applies well enough to theories such as Goldman's and Harris', but not to Gordon's (1995a & b) variation which does not take oneself as a model of others, but rather a generic simulation of what a non-specific person might do, which is then adjusted to make predictions for particular situations and individuals.

microscope this might seem more plausible, though the principle behind both remains the same. This Gibsonian approach does have the advantage of suggesting why commonsense *might* be in competition with theoretical disciplines in some cases but not others. Competition would be possible (though not, perhaps, inevitable) whenever the subject matter is open to be investigated through the senses, and not when it isn't. This is still far from adequate though. While commonsense aeronautics may get a little way off the ground, commonsense electrolysis for example is a complete non-starter as it requires specialised concepts which are not plausibly part of the 'commonsense' repertoire. But the results of electroplating (where a conductive material is plated with a thin layer of another material from a solution, by passing a current through both) are visible to the naked eye, even if understanding the processes behind it is beyond the realms of ordinary 'commonsense'. We might refine things further and say that commonsense is interested in both processes and effects which can be perceived by the naked senses, and the sensible effects of processes which cannot be perceived by the naked senses, but neither processes nor effects which are invisible to the naked senses. This is still far from satisfactory – not least because 'naked senses' is a gross oversimplification – but we don't have the room to pursue things any further.

For the remainder of this section we will assume the broadly Gibsonian view that what is 'commonsense' can be defined relative to the unaided human senses, but bearing in mind we still have no satisfactory answer to the question of how it interacts with science and theoretical reasoning. To the best of my knowledge, in the context of the literature on commonsense psychology – or indeed anywhere else – there isn't one yet to be had.

5.2 Are Ordinary People Aware of Commonsense Psychology?

We mentioned another question which is rarely considered; this is it. Recall that commonsense psychology is almost universally regarded as intuitive and obvious; but is it obvious only to specialists in psychology and philosophy, or also to the lay people who supposedly employ it? On the one hand more-or-less everyone is aware of the fact that we

try to – and succeed in – understanding each other passably well in many cases, and everyone uses the words 'believe' and 'desire' (or synonyms). Whether everyone does so primarily to predict and explain behaviour is – I think – more contentious. At the other end of the spectrum, literally no-one outside of academia (let's say interested amateurs are also academics) thinks that people employ propositional attitudes, if only because they have never heard of them and there is no analogue in common parlance. Perhaps one day 'propositional attitude' will, like 'gene', enter common usage, but it certainly hasn't even begun to yet. Ratcliffe (2007, p. 50-2) conducts suggestive – though statistically insignificant – surveys of undergraduate philosophy students, who almost universally fail to identify prediction and explanation of behaviour in terms of beliefs and desires as an intuitive or likely means of our understanding one another – even when explicitly presented with it as an option. In the vast majority of articles and books on commonsense psychology, the two distinct issues of how people actually understand each other and how philosophers interpret those activities are shamelessly run together, and typically use of the phrase 'commonsense psychology' (or synonyms such as 'folk psychology') equivocates wildly between the two.

Let's look at an example. Fodor writes "if commonsense psychology were to collapse, that would be, beyond comparison, the greatest intellectual catastrophe in the history of our species" (1987, p. xii). His use of 'commonsense psychology' here is clearly equivocal. If it means 'the way people understand each other, *whatever that might be*' he is probably right. If, on the other hand, he means 'the prediction and explanation of behaviour in terms of propositional attitudes' then it is less clear that it would be such a Very Bad Thing. My own view, for what it's worth, is that it would be problematic but not disastrous, as people do not primarily understand each other in those terms – in fact, only really do so when engaged in armchair reflection or contrived contexts such as in certain experiments – but I don't pretend to have argued for this. Yet.

Fodor of course thinks that the two interpretations of 'commonsense psychology' above are one and the same, but this is and must be treated as a substantive philosophical position, not as blindingly obvious or too innocuous to dwell upon. The first interpretation is a straightforward description containing no analysis of what people do whatsoever, whereas

the second involves a considerable degree of analysis. This in no way means that the analysis is incorrect, but to fail to recognise that the second interpretation *is* a substantive claim is to confuse description with analysis. As we have seen before this is easily done – recall that the entire debate over counterexamples to mereological extensionality was based upon just this confusion – but it is a mistake nonetheless. For the sake of clarity, whenever I have or will use the term 'commonsense psychology' I use it in the sense of the substantive analysis of everyday thought in terms of propositional attitudes (unless explicitly stated otherwise).

Equivocating over the two interpretations of 'commonsense psychology' can make the analysis seem more plausible as – whether intentionally or otherwise – it is possible to conduct a 'bait-and-switch' where the contentious analysis is alternated with the anodyne description according to the individual's purposes. Given that most philosophers wrongly view the analysis as uncontroversial, this happens more often than it might.

I once overheard a prominent philosopher of mind – it doesn't really matter who – say in response to someone criticising propositional attitudes as how people think 'Ah, but don't you *believe* that what you are saying is correct…?' seeming to view this as a valid argument. It is nothing of the sort. If by 'believe' the philosopher meant something like 'have a propositional attitude' then his comment straightforwardly begs the question. If he meant 'believe' in a purely descriptive manner then his question was a psychological one about the other person's opinion, and no objection at all to what he was saying. This sort of confusion is endemic in the literature, though not surprisingly there are some very careful and incisive pieces of work which distinguish between the different interpretations of 'commonsense psychology' and deal with them separately. These include amongst others Bermudez (2005) and Ratcliffe (2007), both careful analyses which have done much to inform the discussion here.

Ravenscroft (2004, see also Stich & Ravenscroft) makes a distinction between 'external' and 'internal' theories of commonsense psychology which might be thought to correspond to ordinary people being aware or unaware of it respectively. This is not the case. 'External' commonsense psychology was devised by David Lewis, and is the less popular of the two

approaches. Lewis suggests that it is possible to articulate people's commonsense views in terms of 'platitudes' – everyday observations and sayings which most people would assent to. Homespun wisdom, that sort of thing I suppose. In this it is somewhat similar to G. E. Moore's view of common sense consisting in a mass of propositions we know to be true, without necessarily knowing why or how. Lewis' innovation is to systematically corral all the platitudes together (ruling out the input of weirdos and undesirables), and to treat them as a term-introducing theory. So 'belief' for example is defined by all the platitudes in which it, or one of its derivatives, appears (for technical details see Lewis 1972). Commonsense psychology is the theory made up from all such terms.

'Internal' commonsense psychology, which is standard in the literature, takes the basis of our everyday reasoning to be an internally represented theory of mind.[89] This would be some sort of process or mechanism in the brain which produces the propositional attitudes that supposedly are expressed in our everyday thoughts and speech. So for internal commonsense psychology the process which underlies the psychology itself is implemented in the brain or the mind, at a 'sub-personal' level. By contrast the process which underlies commonsense psychology on the external approach is 'personal' in that people are aware of the platitudes – after all, they use and hear them all the time!

So why doesn't the external versus internal distinction correspond to people being aware versus being unaware? Let's go back for a moment to electroplating. The results are (or at least can be) readily apparent to ordinary observers, while the process remains a mystery to the naked eye and requires a fair degree of analysis to figure out. The analogy with internal commonsense psychology should be obvious, and internal processes are of course not usually available to conscious thought. However, the same analogy holds equally well with external commonsense psychology. The results are obvious enough, and it might be thought that the process – consisting of publicly agreed, commonsense platitudes – is obvious to ordinary people as well. But this is not the case. To start with, by definition everyday platitudes are the sorts of things people recognise and assent to, but that is not to say that

[89] Here for simplicity I use the phrase 'theory of mind' neutrally between the theory theory and the simulation theory. This departs from Ravenscroft's and Stich & Ravenscroft's use of 'internal' in this context, which only applies to the theory theory.

there is in fact enough general assent for them to be systematised in the way Lewis intends. Individual platitudes might be recognised, as sure as eggs is eggs, but it is contentious whether any recognisable body of them exists across groups, much less cultures, nations or ages. This is, of course, an empirical issue. There is no a priori reason I know of to think it either would or would not be possible to form a body of platitudes – it depends entirely on who you ask and what they say. Even if there is such a body to be systematised this highlights that the actual process subserving external commonsense psychology is the implicit structure within a widely spread body of platitudes, *not specific platitudes used and recognised by individuals themselves*, and hence is no more apparent to ordinary people (by which I mean anyone not conversant with the idea of commonsense psychology) than some neurological process, not least because both may very well be false.

People are unaware of commonsense psychology in the interesting sense of being an analysis of how they think, rather than a straightforward description which is essentially a demonstrative ('that!'). In the latter sense only people are aware of commonsense psychology, but this is a quite uninteresting result. Commonsense psychology (in the interesting sense) is a substantive philosophical analysis of everyday thought, and must be treated as such! But the allusions we have made so far to concerns about its adequacy as an analysis are no match for sound and rigorous arguments in its favour. The question is: are there any? And the answer is no.

5.3 Support for Commonsense Psychology

The usual arguments in favour of commonsense psychology fall into four general types:

- It is so widely believed it is beyond question.
- It is so intuitive or obvious it is beyond question.
- It is widely used in cognitive psychology.
- It is supported by specific empirical studies.

We will consider each of these in turn, and see that none of them provides any substantial degree of support for commonsense psychology. The first two argument types are worthless, while the third and fourth provide support which is questionable at best. As we shall see, there is no good reason to think that commonsense psychology is beyond rational dispute, and several good reasons to think it may well be false.

5.3.1 "It is so widely believed it is beyond question"

I haven't ever actually seen this argument in print, but it is one which has been implicit in some conversations I have had regarding commonsense psychology, so I do think it is worth mentioning here. Clearly, this is a fallacy of epic proportions. After all, 5 billion Christians can't be wrong, right? To be fair I do think the premise of this argument supports a slightly different conclusion; most philosophers are (probably) hard-working and serious professionals, so if a large number believe a particular analysis or theory to be true then it is worth taking seriously. That still doesn't make it true though. If once upon a time everyone believed the world was flat, they were all wrong; commonsense psychology may yet turn out to be a 'flat Earth' theory.

5.3.2 "It is so intuitive or obvious it is beyond question"

It doesn't really matter how many, or how few, people believe in something if what you care about is whether it is true. Their reasons for believing in it, however, do matter. It is quite tempting to think this argument holds some weight. After all, we did say that most philosophers are hard-working and serious professionals, ones who have spent years honing their skills through honest toil. So what they take to be *intuitive* about their area of study is highly likely to be correct, and might even be beyond question by those less familiar with the discipline and relevant topics.[90] However, in this specific area – and probably in many

[90] This is a line of response I encountered a couple of times early in my postgraduate studies in connection to expressing doubts about commonsense psychology. It may be that my youthful questions were somewhat fatuous and invited this kind of response, but I doubt it.

others – the exact opposite is far more likely to be the case. Philosophers' intuitions are notoriously corrupt, and the further one delves into subtle and sophisticated analysis of some phenomenon, the harder it is to engage with the phenomenon itself independently of one's own analysis. This is especially true when the exact nature and status of the phenomenon is itself unclear – as we saw with the putative domain of 'commonsense' it isn't too hard to articulate roughly what it is, but it is very tricky to be precise. I don't think it is ever easy to separate description and analysis, and there is a case to be made for claiming it is completely impossible. What is more, the people who are most deeply steeped in analysis aren't necessarily the best candidates for the job.

I want to stress that this is not in any way an anti-philosophical position, nor a blanket criticism of any group of philosophers themselves. The point is that it is a fallacy to suppose that those who are most heavily engaged in analysing a phenomenon are likely to be in a privileged position to describe that phenomenon independently of their own analysis. With exceptions, overall they may even be less likely to be able to do so than someone ignorant of the analysis in the first place. Against this point it could be argued that an expert theorist will have the advantage of being more able to pick out the relevant features of the phenomenon[91]; with regards to commonsense psychology, it may be that philosophers have successfully picked out the central role of propositional attitudes, and explanation and prediction in their terms, while lay observers are likely to settle upon all manner of features which may or may not be of central importance.

There may well be something to this response – but notice its effect is to engage in the question of whether philosophers' insights or intuitions should be accepted in this specific case, not to defend their blithe acceptance regardless of what they might claim. Bald intuitions are worthless as evidence, doubly so when they are shaped by a well entrenched and highly refined theoretical perspective. Perhaps, then, the origins of these intuitions might lend some support? They are widely cited as stemming from two discussions – Lewis' systematising of platitudes, and Sellars' just-so story of our 'Rylean ancestors'. Neither of

[91] This is, for example, a feature of the 'constructivist' paradigm in educational theory. See e.g. Grimmitt.

them provide any good reason to believe commonsense psychology should be accepted on intuitive grounds.

Lewis' technique for producing a term-introducing theory out of commonsense platitudes may be empirically impossible, as we have seen (is there a body of platitudes to systematise? In fact, do people think largely in terms of platitudes at all?), but even if possible it in no way shows that people do think or act according to such a theory. All it demonstrates is that such a theory can be constructed. I am reminded of Tuomela's book *The Philosophy of Social Practices* in which he analyses these slippery phenomena in terms of 'collective acceptance' – roughly, combining mutual belief with mutual recognition of belief and mutual recognition of that recognition. The section relevant to us here is chapter 7, where he attempts a systematic analysis of the forgoing theory, to no apparent benefit at all. As any undergraduate studying the dark art of elementary logic knows, it is laughably easy to translate statements or ideas into a systematic language (with propositional or predicate logic being the usual instruments of choice) – all you have to do is make some marks on paper! Better yet, save some ink by doing it in your head. The real issue is whether it has been done in a way which is of some identifiable benefit or usefulness (or as the students say, did they get it right?). I must admit to being nonplussed as to why Tuomela seems to view it a merit that his theory can be rendered in a formal system. After all, anything can be formalised in any way you like: how much more impressive it would be to be able to produce one which can't! This might seem flippant, so let's take another example. At Durham University I contributed to teaching introductory logic for a few years, primarily using semantic 'trees' rather than the more traditional natural deduction. This method of systematising the semantic relations between proposition types is worthwhile because it delivers correct results (within the confines of first-order predicate logic, and subject to some qualification to avoid infinitely long 'trees'). The tree method's appropriateness is guaranteed because its rules are provably sound and complete with regards to the logical system it is used with. Anyone can dream up some rules or a technique – the real issue is whether it works.

To indulge in an anecdote, this point became clear to me at the tender age of nine when I invented what I took to be a new mathematical method of solving problems, which I enthusiastically applied to everything I possibly could. It was a form of iteration, though I didn't know it at the time, and I was rather impressed that it often delivered answers which were close to the correct one while other means failed me entirely. Over time, however, I came to realise that I rarely got the sums exactly right. Eventually I was forced to admit that my wonderful new method wasn't much use for the tasks at hand and went back to more traditional methods, learning to apply them properly.

Back to Lewis. His supporters will no doubt respond that the usefulness of his systematization can be, and has been, demonstrated by its employment in numerous arguments and articles. I don't disagree that it has been put to use, but whether it has been beneficial in attempting to uncover how people actually think is precisely the issue at question in this section. My point here is simply that the Lewisian method of systematising platitudes doesn't provide any support for the view that commonsense psychology is intuitive or obviously beyond question. It might work well enough to be worthwhile (though considerations raised in this section suggest otherwise), or like my youthful iterations it may prove to be limited in value and best avoided. Either way, it doesn't make commonsense psychology remotely obvious or intuitive.

What about Sellars' just-so story? As part of an attack on the received Cartesian view that there is some sort of infallible first-person access to mental contents (the 'myth of the given'), Sellars describes a fictional group of ancestral behaviourists who can characterise behaviour semantically. Over time they realise that behaviour can be best explained and predicted by hypothesising internal acts of speech which they call 'thoughts', and from that basis develop a theory of mind including experiences, impressions and various other features. This story is generally credited as the origins of the theory theory of mind (see e.g. Stich and Ravenscroft, Gordon 2000), but offers no support for the view that commonsense psychology (or any more specific variety such as the theory theory) is intuitive or obvious. Sellars himself is clear that it is just a story – he in no way suggests that this re-casting of the development of functionalism as human history is veridical.

Perhaps, then, a third reason why commonsense psychology is seen as intuitive – it is congruent with functionalism in the philosophy of mind, which is by far the dominant view of how the mind works. Functionalism, to a fair approximation, takes thought to consist of series of discrete mental states which are characterised by their causes and their effects rather than any intrinsic properties. Commonsense psychology's propositional attitudes are somewhat similarly discrete states which are characterised in part by their propositional content, and in part by their 'direction of fit': in the jargon belief has a world-to-mind direction of fit, while desire has a mind-to-world fit.

Functionalism doesn't entail commonsense psychology for two reasons. One, commonsense psychology requires mental states to have some intrinsic property to differentiate them according to their contents (e.g. for RTM an appropriate internal structure). Two, functionalism is consistent with alternative descriptions of how we think and interact. Conceptual role semantics (CRS) is a case in point. This is a variation on the Wittgesteinian view that meaning lies in usage, developed by Block, Harman and others. Weak versions of CRS claim that a state's having mental content depends upon having a functional role, while strong versions claim that a state's specific content depends upon its particular functional role (Block). This characterisation of mental content is essentially holistic, and under at least some variations explicitly at odds with the compositional semantics of the RTMs discussed in this thesis. Nevertheless, it is a functionalist theory.

Commonsense psychology also does not entail functionalism; plausibly it requires mental states in the form of propositional attitudes to have some intrinsic properties, while functionalism claims that they do not. The two are certainly congruent in many respects (enough to motivate Bermudez' view that at least some RTMs may usefully be thought of as a development of functionalism to include structured mental states), but neither one requires the other.

To summarise: philosopher's intuitions are influenced heavily by their theoretical commitments, and hence difficult to assess clearly. No reason typically given to support intuitions regarding the obviousness of commonsense psychology in any way entails that the theory is true.

5.3.3 “It is widely used in cognitive psychology”

This argument deserves taking more seriously, especially given our emphasis on the value of empirical data in formulating descriptive ontologies. On one reading, the argument works well: widespread use of a concept in cognitive psychology can demonstrate that the concept is useful or even essential in characterising the workings of the mind, hence needs to be part of our philosophical vocabulary. However, there is an implicit premise which needs to be spelled out. This is that the use is essential and cannot be dropped without loss of meaning or significance (including being recast in terms of some other concept which is already recognised). Despite the widespread use of ‘commonsense psychology’ in cognitive psychology, it can usually be eliminated without loss; hence the argument is inductively weak at best and an outright failure at worst.

It may be that some uses of ‘commonsense psychology’ in cognitive science do satisfy the enthymematic premise, and do provide some support for the theory. I know of none. Of course, I cannot pretend to have read more than a tiny fraction of all that cognitive psychology has to offer, and so can only comment on what I have seen, asking the reader to judge whether my experiences concur with his or her own. Overall, I am certain there is no *widespread* support from cognitive science for commonsense psychology, which is what the argument depends upon. But how can this possibly be true when talk of ‘commonsense psychology’, especially in the guise of the phrase ‘theory of mind’, is rampant in much cognitive psychology?

The issue is that use of these phrases which makes any claim at all about their nature is much rarer, and in the absence of any detail of their use being relevant to the main topic of the piece of research it seems likely that they are being used as a placeholder for ‘whatever system people use to understand each other’. Recall earlier in this section we discussed the unfortunate equivocation between this use of ‘commonsense psychology’ and its use as a specific theory of how we think – to imagine that simply using the phrase ‘commonsense psychology’ or any cognate notion is to commit to a particular theory is to fall foul of just

this point.[92] Of course, if any article makes specific use of prediction and explanation in terms of propositional attitudes – and bases its conclusions or results upon them, not just mentions them in passing – then it should be counted as evidence in favour of commonsense psychology, but this is beyond rare.

Non-committal passing references to commonsense psychology (or theory of mind) may be ubiquitous in cognitive psychology, and brief descriptions or glosses on its meaning are common, but both can be eliminated without loss of meaning or significance to the work in question. Ineliminable references are rare at best, and I know of no definite examples; they can provide very limited support for commonsense psychology as an analysis of how people think.

5.3.4 "It is supported by specific empirical studies"

A number of specific empirical studies are often cited in relation to commonsense psychology, in particular Wimmer and Perner's False Belief task and variants upon it. In fact, these studies are so riddled with questionable assumptions they do not provide any clear, direct support for commonsense psychology at all.

The false belief task is set up to detect at what age children are able to attribute specific mistaken beliefs to others; in the original version children watch a puppet called Maxi put chocolate in a blue cassette-box 'cupboard'. Maxi's mother then removes the chocolate in Maxi's absence and places it in a green 'cupboard'. Children are asked to point to the cupboard where Maxi will look. Wimmer and Perner report that most six to nine year olds pass the test by pointing to the original cupboard, four to five year olds tend to pass when asked to think carefully, and three to four year olds tend to fail regardless of whether they are given such prompts. This task, including variations upon it, is frequently regarded as a 'litmus test' for the development of commonsense psychology (see e.g. Frith & Happe, p.3) at around four years of age.

[92] The same point applied earlier to equivocal uses of 'mental representation' and 'mereology'. Philosophers, beware equivocation!

The rationale is this: commonsense psychology requires an accurate concept of belief to function at all (how could someone predict and explain accurately in terms of beliefs, if they don't know how beliefs work?). One of the key features of the concept is that beliefs can be mistaken; the false belief task seems to show that before 4 years of age children do not appreciate this. Hence, their conception of belief is badly flawed. It follows that their ability to understand others – in the guide of predicting and explaining behaviour – is likewise severely limited, and can only really function properly once there is an accurate concept of belief in place. This is not to say that children under 4 cannot understand other people or the world around them, rather that the methods they use are distinctly impoverished compared to full-blown commonsense psychology, which is only possible once false beliefs can be attributed to people.

Although *prima facie* quite reasonable sounding, there are two significant flaws in the suggestion that the false belief task provides evidence to support commonsense psychology. The first is that in order to even understand the false belief task, children need to have quite a sophisticated grasp of the social situation. Not only do they need to understand what the experimenter is asking them to do (Gallagher 2001, p.99), they also need to recognise that the puppets and other objects are stand-ins for agents and furniture, and be able to follow a pretend narrative involving them (Ratcliffe 2007, p.53). This is not in itself too problematic – after all, no advocate of commonsense psychology claims children lack *all* social understanding until they can pass the test, just anything resembling commonsense psychology. The interesting question is whether, as Gallagher (2001) claims, this 'primary intersubjectivity' which precedes the supposed onset of commonsense psychology is also primary in the sense that it remains at the forefront of how we understand each other throughout life; this question is also developed by Ratcliffe (2007), especially chapters 4 and 6. Although partly based upon empirical studies, Gallagher and Ratcliffe's arguments also draw heavily on phenomenology and as such lie outside the scope of our discussion. Nevertheless, I would like to make clear that I take their method of combining empirical data with phenomenology to be appropriate and valuable, and also complementary to both the methods used and conclusions drawn in this thesis. We will return to the issue of 'primary intersubjectivity' in discussing the core knowledge hypothesis and mirror neurons below, and conclude that there is plenty of empirical evidence (with or without

phenomenology) to support the view that infants possess a degree of understanding of the world around them – including other people – which is continuous with and forms the basis of later cognitive achievements.

The second flaw is that the entire experimental design is riddled with assumptions which favour commonsense psychology's view of its results. The false belief task assumes that interpersonal understanding is a matter of predicting future actions based upon overt – and relatively crude – physical behaviour. Use of puppets abstracts away from all subtleties in body language, gaze and shared attention, as well as from the fact that most interactions young children engage in involve people familiar to them, with the children as active participants not passive observers. Small wonder that the results of such a test should be congruent with commonsense psychology, when many features of just that theory are built into the experimental conditions themselves! Commonsense psychology emphasises observational understanding, as does the false belief task; commonsense psychology takes one of the main purposes of cognition to be prediction, as does the task; and commonsense psychology emphasises the inference of mental states such as belief on the basis of overt behaviour, as does the task. As Ratcliffe neatly puts it, 'The assumption that adults understand each other by attributing internal propositional attitudes was written into these studies from the start' (2007, p.55).

Given all this there is absolutely no way that the false belief task can be clearly said to provide any clear or direct empirical evidence in favour of commonsense psychology. The two are certainly complementary, but this is only because of shared fundamental assumptions about their subject matter. To be fair, we have not canvassed variations of the task, but these likewise share the same assumptions; see Wellman *et al*. for a thorough meta-analysis of the various false belief-style experiments, which demonstrates an overall consistency in methods and results.

The false belief task plausibly does demonstrate a genuine cognitive achievement, a significant conceptual change which occurs around four years and contributes to our ability to understand others. However, if commonsense psychology is correct then this milestone heralds an enormous cognitive leap – suddenly being able to start using the system by

which *all* older children and adults understand each other! We would naturally expect to see a rather dramatic step up from an impoverished interpersonal understanding before four years or so, to a rather sophisticated one very shortly afterwards. Nothing could be further from the truth. In the following section we will examine empirical data which demonstrates a continuous development of cognitive capabilities – including understanding of other agents and their actions – from early infancy. Passing the false belief task is one small cognitive milestone amongst many others throughout life, not a litmus test for the ability to begin understanding other people.

5.4 Core Knowledge, Mirror Neurons, Body Image and Body Schema

We have seen that there is very limited empirical support for commonsense psychology, if any at all. The aim of this section is to show that there is an abundance of empirical data which directly undermines it. In particular, to show that:

1. The sudden development of the ability to understand others (by developing a concept of belief) at around four years is highly implausible.
2. Describing how the mind works requires concepts which cannot be reduced to propositional attitudes or their uses.

If belief is the cornerstone of our ability to understand the world around us, and young children have a severely deficient conception of belief, it follows that they have significant difficulty in understanding the world around them. The moment the correct concept is learnt, there ought to be a colossal leap in the ability to interact with others and understand one's environment. Enormous vistas of opportunity to predict and explain, previously denied, would suddenly spring forth. The best empirical evidence currently available suggests that both points are abjectly false. In contrast humans and non-human primates possess a set of 'core' competencies or systems which are shown in humans from very early infancy, and which persist as the basis of more sophisticated conceptual abilities throughout childhood and early adult life. This 'core knowledge' hypothesis predicts, and is confirmed

by, the considerable ability young infants show in understanding salient features of the world around them.

Empirical data drawn from clinical neurology about the functioning of the mind demonstrates the existence of two distinct systems of attitudes towards our own bodies, called the body image and body schema. Neither concept can be adequately captured by commonsense psychology, showing that it is an impoverished account of the workings of the mind.

5.4.1 The Core Knowledge Hypothesis

Advocates of the hypothesis, in particular its leading light Elisabeth Spelke, reject the dichotomy between viewing the mind either as consisting in a single highly flexible system (e.g. Humean associationism), or as a large number of highly specialised mechanisms evolved to address specific problems (to use the jargon, either modules or heuristic devices. The best known representative of this view is probably Pinker 2002). Rather, there are held to be a number of 'core' systems which are present from early infancy and persist though into adulthood whilst more conceptually sophisticated abilities are built upon and around them. These core systems are specific to particular domains of objects, certain types of tasks, operating largely independently of one another (Spelke 2000, p.1233), and as such the hypothesis is probably best viewed as a synthesis of the two dichotomous extremes. Core systems are more numerous and less flexible than the first extreme's single central system, but far less numerous or specialised than the second's highly specialised mechanisms.

There is a very substantial quantity of empirical data derived from over 20 years of studies to support the core knowledge hypothesis. To the best of my knowledge there has been no real attempt to relate their findings to the philosophy of mind.[93] This isn't a *real* attempt either, since our discussion will be quite brief, but it will pick up on the major themes of the

[93] I am excluding Xu (1997) and Wiggins (1997), since they focus only on one small aspect of the hypothesis, how to gloss the conception of physical object apparently at play.

hypothesis, and hopefully go some way to demonstrating how impoverished the view of the mind coming from commonsense psychology is.

The core knowledge hypothesis identifies at least four systems (Spelke 2004, Spelke & Kinzler): those which deal with knowledge of inanimate objects and their physical interactions, agents and their goal-oriented actions, approximate numerical magnitudes and their addition, subtraction and ordering, and using geometric features of the environment to orient oneself. Spelke & Kinzler identify a fifth system, to differentiate membership of social groups, but this has rather less empirical support so far.[94]

Each of the four core systems has been investigated using variations on the preferential looking methodology – subjects (usually infants, but also young children, adults and non-human primates) are 'habituated' to a display by being repeatedly exposed to it until the time they spend looking reduces to a level deemed suitable, such as 50% of the time spent at the beginning of the process (Spelke 1990, p.32). At this point the subjects are shown a different display (or in the case of control groups, more of the same) and the time spent looking is measured and compared with the previous level. By using specially trained observers, the times can be measured to a high degree of accuracy, and an increased looking time is taken to indicate surprise or interest at the difference(s) in the new display. A similar or identical looking time either indicates a failure to notice, or a lack of interest in the change(s). The beauty of this method is that it can be used on any human or animal capable of staying reasonably still and looking at something for a fairly short period of time – it is routinely used to test infants as young as three months. This methodology is regarded as highly robust amongst cognitive psychologists, and I know of no significant attempts to undermine or challenge it.

[94] An interesting question is whether 'core' systems are conceptual or not. I am inclined to think that they are: there is a core set of relatively simplistic conceptions which have a largely biological basis, and which are refined and developed throughout life. However, I cannot think of any particularly strong argument to show that the abilities shown here need to be thought of as being conceptual. The best, perhaps, is that the hypothesis takes there to be a continuous process of development, with later systems definitely being conceptual, but this is by no means conclusive. A non-conceptual process could plausibly become conceptual as it develops, although I wouldn't much like to speculate on how.

5.4.2 Inanimate Objects

Of the four systems, this has attracted the most attention and research. I don't see a need to repeat the detailed summaries given by Spelke (2000, 2004) and Spelke & Kinzler, but instead will pick out the main points relevant to our purposes. Numerous experiments have tested the reactions of infants to variations in the behaviour and properties of inanimate objects, finding that even from a few months old infants have quite clear expectations about them. Inanimate objects are expected to show

> cohesion (objects move as connected and bounded wholes), continuity (objects move on connected, unobstructed paths), and contact (objects do not interact at a distance).
>
> Spelke & Kinzler, p.89

Young infants show surprise or interest through preferential looking when objects appear to move through each others' surfaces or boundaries or undergo quantitative changes. They also have clear expectations about when objects should move or come to rest (when they are struck by another moving object, or come into contact with a stationary one, respectively), and when they should remain stationary (no action at a distance).

Infants are also insensitive to many qualitative differences – unlike for adults and older children, their judgments of objects' unity showed no variations based upon colour or texture, only continuities in surface dimensions, movements and arrangement (Spelke 1990, pp.36-8). Strikingly, Simon *et al.* demonstrated that although infants are sensitive to variations in the number of toys revealed after being placed behind a screen (Wynn 1992), they did not react to qualitative changes such as size, shape and colour produced by replacing the toys with different ones altogether.

Furthermore, the results have been shown to be multi-modal, with haptic (touch-based) variations showing similar results in that infants were more likely to judge two rings to be part of a single object where they rotated in unison and/or rotated about the same fixed point (see e.g. Streri *et al*. 2004). There is also evidence to show that they may be domain-specific: Huntley-Fenner *et al*. repeated Wynn's experiment with minor variations using

sand piles, as did Chiang & Wynn with piles of blocks, both failing to reproduce her findings for masses rather than discrete unified objects. Rosenberg & Carey found that 8-month-old infants fail to discriminate one from two sand piles, whilst successfully discriminating solid objects which resemble sand piles. This is clear evidence that infants as young as five months are able to discern objects which possess a greater degree of unity than a mere heap. It has been suggested by Xu (1997) that despite their tender age they are employing a 'dummy sortal' concept of *physical object*, meaning roughly 'bounded, coherent, 3D object that moves as a whole', though see Wiggins (1997) for a strong rebuttal of Xu's proposal.

In light of the considerable weight of empirical evidence there can be no reasonable doubt that from an extremely young age infants are capable of effectively tracking material objects in their environment by some of their properties – interestingly enough, almost exactly those identified by Locke as primary qualities (the exception being figure) – whilst they remain apparently oblivious to others, corresponding at least roughly to the secondary qualities.

5.4.3 Agents and Actions

Preferential looking methods have also been widely used to test whether infants respond differently to humans versus inanimate objects, finding a marked difference arising by six months (see e.g. Johnson for a review). In a recent study Molina *et al* use this method to test six-month-old infants' responses to situations where human agents either talk to or grasped and manipulated other agents or inanimate objects. The infants looked longer when either an inanimate object moved after being talked to, or one person grabbed hold of and moved another. Further experiments showed that the details of the movement or objects involved was irrelevant, and that the same distinction was not made at four months.

Considerably more has been demonstrated by numerous other studies: unlike material objects, agents are not expected to be cohesive or move continuously, but their movements are copied by the infants; they are expected to act in ways which are goal-directed and

efficient; they act contingently based upon other actions and reciprocally; and whilst agents need not have perceptible faces, where they do their direction of gaze is used as a clue to their actions (for references see Spelke and Kinzler, p.90).

Infants display a considerable degree of intersubjectivity from as little as six months of age – by this point they can not only differentiate between animate and inanimate objects, but have quite clear expectations about how the two should act, and respond to them quite differently. And this is at an age where they are unable to speak, probably cannot crawl, and – if the parents follow current NHS guidelines – are only just beginning to attempt solid food. Even if the false belief tasks do uncover a significant developmental milestone in intersubjectivity, it is abundantly clear that it is a development built upon considerable prior ability, in line with the core knowledge hypothesis.

Recent research into mirror neurons has also shed some light on early intersubjectivity. As mentioned above, infants have a tendency to imitate the movements of other agents, and this is now widely attributed to the action of so-called mirror neurons. Their existence was originally demonstrated in macaque monkeys using functional magnetic resonance imaging of the ventral premotor cortex (area F5). The same group of premotor neurons which discharge when the monkeys perform a goal-related hand action such as grasping an object, also fire when the monkey watches another monkey or human do the same (Gallese *et al*. 1996).

Many studies have found an apparent homologue in the human brain, composed of the rostral part of the inferior parietal lobule and the caudal sector (pars opercularis) of the inferior frontal gyrus (IFG), plus the adjacent part of the premotor cortex (for further details and references see e.g. Gallese *et al* 2004, Rizzolatti & Craighero). The human mirror neuron system does display some differences – there need not be an object present for observers' neurons to discharge when watching an action, and the neurons also discharge when watching actions which are not obviously goal-oriented. Mirror neurons are limited not only to observing actions, but also emotions including disgust (for a summary of relevant studies see Gallese *et al.* 2004, pp.397 - 400).

Not surprisingly, the discovery of mirror neurons has spawned massive interest and a correspondingly huge literature, which we will not attempt to survey here. Many findings are highly congruent with studies on imitation in humans and primates, in particular relevant differences between the two. Byrne & Russon differentiate between 'action-level' imitation where a detailed and linear sequence of actions is copied, and 'program-level' imitation where the overall structure of the activity is imitated but not necessarily the precise actions involved. Studying food preparation techniques of wild mountain gorillas and the imitation methods used by orangutans being rehabilitated to live in the wild, he has found that apes use primarily program-level imitation in skill learning. In contrast, human infants are much more likely to use action-level imitation. On the face of it, this might suggest greater intelligence on the part of the apes, but there are other possibilities. It may be that action-level imitation plays a social role such as demonstrating group membership (something like a secret handshake, I suppose), or helps understand an overall goal or technique (on the assumption of rationality in the agent observed, there must be a reason why he does it the hard way). The latter is supported by Gergely *et al.* who observe that 14-month-old human infants are more likely to imitate an unusual action (switching a light on with one's head, or using an odd tool) when the agent appears to do so freely than when the agents appears to have no choice. Non-human primates imitate in the same way regardless (Buttelmann *et al*.).

What is clear is that mirror neurons' existence is highly suggestive of a shared experiential basis for action and emotion which precedes the supposed development of a theory of mind at around three to four years. Some have taken their discovery to support the simulation theory over the theory theory, e.g. Gallese and Goldman, since they make a direct link between our own actions and those of others, and could conceivably be an evolutionary precursor to more sophisticated simulation, whether 'putting oneself in another's shoes' or developing a generic model of human action. The problem with this view is that mirror neurons seem to collapse the distinction between perception and action entirely – from early infanthood the two go hand-in-hand. A basic aim of both simulation theory and theory theory is to explain how we can go from a solipsistic understanding of ourselves to be able to explain and predict the actions and thoughts of others; how can we bridge the gulf from

the subjective to the intersubjective? Mirror neurons provide direct empirical evidence which questions whether this gulf even exists in the first place.

From as early as six months, infants differentiate between agents and objects with clearly distinct expectations about the behaviour of both. Mirror neurons show that there is a neurophysiological basis for this distinction (admittedly in slightly older infants, but still at an exceptionally young age), and that the epistemological gulf between self and other is rather less pronounced than theory theory and simulation theory suggest.

5.4.4 Numerical Magnitudes and Geometrical Orientation

These two systems are less directly related to commonsense psychology, but nevertheless help illustrate an important point about the core knowledge hypothesis. Both illustrate well why the systems in place in infants need to be regarded as continuous with later more sophisticated systems.

Using preferential looking, Xu & Spelke (2000a, 2000b) showed that 6-month-old infants discriminate between displays containing 8 or 16 dots, or 16 or 32 dots, by looking longer at a novel numerosity, but do not discriminate between changes in size, density or position. They also did not discriminate 8 from 12 dots, or 16 from 24 dots, showing sensitivity to a ratio of 1:2 but not 3:2. Spelke (2000) also references an unpublished manuscript by Lipton & Spelke, which claims to show that the ability to discriminate relative numerosity is multimodal, as it applies to sounds as well as visual displays. Interestingly, the 1:2 ratio only applies to larger numbers, not to 1 versus 2 dots (Xu & Spelke 2000a), suggesting two distinct systems present in young infants for dealing with numbers (Spelke 2000, p.1237) – one for exact, small numbers (as tested in Wynn), and another for larger numbers.

These findings nicely illustrate partly why 'core' systems should be viewed as the basis of later developments rather than being superseded entirely. Barth *et al.* tested adults using a similar methodology to Xu & Spelke, with set sizes from 10 to 70 and ratios from 1:3 to 6:7.

Adults showed an equivalent ability to differentiate between approximate large numbers of dots to infants, but much more accurately: they easily succeeded with 3:2 and showed a better than chance result when comparing 60 to 70. Although adults' abilities in this area are superior, there is no reason to think there is any *different system* at work, just the same one which has been refined to increase accuracy and cope with larger sizes.

A similar finding applies to discrimination of absolute number in physical objects. Infants discriminate between 1 and 2 objects readily, whilst ignoring qualitative features such as colour and texture. Adults and older children still differentiate between objects quantitatively, extending their range to higher numbers, and add to this the ability to differentiate based upon qualitative features of the objects as well (their secondary qualities, if you like). At no point do the older children or adults lose the abilities of early infancy – they simply extend and refine them. This is the heart of the core knowledge hypothesis, one which is well corroborated by the facts as they stand.

The orientation of infants and adults likewise supports the hypothesis. Gouteux & Spelke survey ten experiments which show that 3-4 year old children do not orient themselves based geometric features of the positioning of nearby identical objects, or non-geometric features of their layout. They did orient themselves based upon geometric features of the walls. In contrast, adults used both geometric and non-geometric features of both the environment and objects within it to orient themselves. This, again, shows the development and extension of a 'core' system which nevertheless remains in place through later life.

5.4.5 Core Knowledge: Key Findings

There is a vast experimental literature (which we have scarcely skimmed the surface of here) to confirm the hypothesis. For our purposes the hypothesis has four main findings:

- Infants as young as three months are able to discriminate between different types of things in the world around them, based upon their salient features.
- From six months, infants discriminate between inanimate objects and agents, with clearly distinct expectations about the behaviour of both. Mirror neurons provide a neurological basis for this early grasp of intersubjectivity.
- There are at least four systems, each of which is highly domain-specific.
- The development of more sophisticated means of understanding the world around us and each other is an ongoing process of extending and refining these early systems, not a case of sudden or wholesale replacement.

Before discussing the implications of these findings for commonsense psychology, we will turn to a second set of empirical findings – those used by Gallagher to argue for a distinction between body image and body schema. Our purpose will be to show that the vocabulary of propositional attitudes is inadequate to capture the workings of the mind. Unlike most uses of 'commonsense psychology' in the cognitive literature, there are other terms which are of great importance for our understanding of the mind and each other which cannot be eliminated (or reduced to propositional attitudes) without significant loss. Body image and body schema are two of them.

5.5 Body Image and Body Schema

Shaun Gallagher (2005a, 2005b) has argued a proper understanding of the interplay between body and mind – including how each influences the other – requires the development and use of two key concepts: body image and body schema.[95] The body image is a 'system of (sometimes conscious) perceptions, attitudes, and beliefs pertaining to one's

[95] Gallagher did not invent the terms, which have been in use in various disciplines for decades. However, previous uses have varied enormously, with no clear or consistent meaning for either term. Gallagher's project is to both give a 'clear definition' of the two (see Ramsey 2006), and motivate their widespread use.

own body' (2005b p.234; see also 2005a p.24) while the body schema is a nonconscious system of processes that constantly regulate posture and movement' (*ibid.*). Gallagher draws evidence for both from a range of neurological and other clinical studies, arguing convincingly that a double dissociation is shown between them. Some patients show impairment in body schema without body image, whilst others show impairment in body image but not body schema.

The principle evidence that a body schema can be impaired without loss of body image comes from the case of Ian Waterman – often referred to as IW – who suffered from damage to his sensory nerve fibers called deafferentation, resulting in losing all sense of movement and position as well as cutaneous touch from the neck down. The initial effect was paralysis; IW gradually became able to sit upright and move his limbs in a controlled way, but only when thinking about and looking at his body. As soon as he looks away or loses concentration his muscles relax, and unless supported he then collapses altogether. It seems clear that IW lost his normal ability to regulate his posture and movement unconsciously, and developed a passable substitute through conscious effort. In Gallagher's terms, he learnt to employ his (largely) intact body image in a novel way to compensate for the loss of his body schema.

Conversely, Gallagher argues, patients suffering from unilateral neglect display an intact body schema alongside a reduced body image. Unilateral neglect is usually caused by a stroke affecting the right cerebral hemisphere, resulting in a lack of attention to the left side of the body. This can manifest in many ways, including failing to dress the left part of one's body and failing to recognise its existence in conversation. Interestingly enough, though, sufferers will use both hands in activities they engage in without conscious reflection such as catching a ball thrown to them or tying a knot.

Gallagher also draws upon neonate imitation, phantom limbs and gesturing as evidence to further support his distinction, but these two examples are sufficient to illustrate a clear double dissociation. They provide both empirical evidence for the distinction between body image and body schema, and an illustration of the concepts' value. For my part, I am quite persuaded of the value of the distinction, but others may be less so. In particular, while the

distinction may be useful, could it not be eliminated by being expressed in terms of some other concepts already recognised?

It would be particularly germane to advocates of commonsense psychology if the two concepts could be recast in terms of belief, or other propositional attitudes, but this simply is not possible. The one term 'belief' cannot reasonably be univocally applied to the range of attitudes shown just by IW and unilateral neglect sufferers, as they encompass explicit conscious thought and attention, and omissions thereof, and also unconscious attitudes and abilities and their lack. If you throw a ball to someone with unilateral neglect and they catch it with two hands, did they believe it was coming towards them and desire to catch it with two hands? Perhaps, but not in the same sense that they might believe they only have a right side of their body and desire to get dressed – otherwise the net result is that sufferers are irrational or contrary, which they patently are not.

Perhaps we should distinguish between conscious and unconscious belief as being distinct and potentially contradictory. That way we could say, for example, unilateral neglect sufferers believe unconsciously they have two sides to their body, whilst consciously believing otherwise. But, Freudian overtones aside, this is still not satisfactory. The neglect shown is by omission, not commission, so there is no real prospect of being able to identify any determinate mental content for a conscious belief relating to the neglected side of the body. What do you believe about something you don't think about? Purely unconscious beliefs would be no better – since they are unconscious they can only be articulated by overt behaviour, which is no respecter of the determinacy of propositional content. The distinction between body image and body schema stands, with no identifiable prospect of being reduced to beliefs or other propositional attitudes.[96] Given time and space which is not available here, it would be interesting to find out just how many other empirically derived concepts this applies to.

[96] Some commentators even swing the other way, suggesting *more* distinctions are needed rather than fewer in this area. Stamenov argues in favour of treating body image and schema as plurivocal terms, against Gallagher's univocal definitions, and de Vignemont manages to detect four core notions of the body implicit in Gallagher 2005a.

Unlike core knowledge and mirror neurons, the literature on body image and body schema is small, and to a large extent revolves around Gallagher's work as well as that of Jonathan Cole, a clinical neurophysiologist and some-time collaborator of Gallagher's. Nevertheless, some interesting applications of the distinction have been suggested already. Mishara argues that it is critical to establish a core deficit in schizophrenics' bodily experiences, and using a range of neuroscientific sources argues that schizophrenics compensate for a lack of body image by effortfully constructing incomplete proxies from their body schemas, which are themselves disrupted. In a somewhat similar vein, Sauvagnat argues by analogy of symptoms for continuity between childhood autism and schizophrenic-like psychoses, with sleep disorders, repetitive movements, self-harm and other symptoms arising from lack of structure in the body image. Cole also discusses the impact – both positive and negative – on the body image arising from physical sensations and their absence. Stamenov relates the distinction to mirror neurons, arguing that the neurons' primary function in infancy is to extract a basic body image from an innate body schema. The field of applying the body image versus schema distinction, and exploring its relation to other cognitive concepts, is wide open and promises to deliver many valuable results in time.

5.6 Back to Commonsense Psychology

Preferential looking studies have shown that young infants possess significant competence in understanding the world around them, with specific expectations about the behaviour of both inanimate objects and agents. This suggests a substantial degree of intersubjective understanding is biologically innate and appears in early infancy, not at around four years. I believe that there is ample evidence (and much more than described here) to demonstrate that the core knowledge hypothesis is true in essence, and will remain so even if some details vary as it is further explored. As such, the ability to ascribe false beliefs in a highly circumscribed and artificial experimental context is no doubt a significant step in intellectual development and understanding of others, but no more than one step amongst myriad others in a long and continuous process of development beginning in early infancy at the

latest – perhaps much earlier. The importance of the false belief task in understanding the origins of intersubjectivity has been blown out of all proportion.

There are also at least two concepts concerning the mind which are directly supported by empirical data, yet cannot be integrated into commonsense psychology's propositional attitudes and their uses, suggesting deficiencies in the theory's vocabulary. There is, I think, an obvious objection here though: body image and body schema are all very well as scientific concepts, but they arise from unusual neurological disorders and so we should not expect them to be part of the way in which regular people think about each other. This is a fair point, and I don't expect that the terms to be widely bandied about over dinner any time soon (though how many people said that about 'gene'?); Gallagher himself observes that 'in our everyday behavior these things are not so unambiguous' (quoted in Ramsey), meaning the two systems probably do not so clearly dissociate except in cases of neurological trauma. Nevertheless, the point stands. Armed with knowledge of the body image and body schema distinction, and sufficient neurological evidence, we can explain and predict the unusual behaviour of a stroke victim suffering unilateral neglect. Using propositional attitudes alone, the best we can do is attribute conflicting beliefs to explain the striking differences in behaviour described above; a crude and utterly inadequate explanation.

The real problem with commonsense psychology is this: it is a poor description of how people actually think and understand each other and the world around them. An adequate account of intersubjectivity and the ways in which we understand the world around us, each other and ourselves cannot possibly be couched purely in terms of propositional attitudes and their uses: this is at once too sophisticated and too crude. Too sophisticated in the sense that it demands too much of its practitioners: as the false belief task is supposed to show, accurate belief ascription only takes place around four years of age, yet a high level of intersubjectivity is shown in infants by the age of fourteen months. Too crude because, as Gallagher's use of the body image versus body schema distinction shows, there are important cognitive distinctions which it ought to be possible to include in commonsense psychology, but cannot be.

Commonsense psychology's vocabulary of propositional attitudes and their functions is inspired, though not entailed, by functionalism and is simply too restrictive to describe the mind and its functionings at a level of accuracy in line with current knowledge of its workings. It is Newspeak, in the Orwellian sense of an official language whose reduced ersatz vocabulary guides and limits what can and cannot be expressed or thought. Any behaviour or thought process *can* be described in terms of beliefs and desires, and that description *can* be used to produce an explanation or a prediction of behaviour. The real issue is how useful this method is, and there is abundant empirical evidence to suggest that it isn't very useful at all. Here I depart from Ratcliffe, who says commonsense psychology doesn't work at all. It works, but works really quite badly.[97] Of course, unlike Orwell's Newspeak, commonsense psychology was never intended as an intellectual restraint – rather, it is supposed to be a clear and systematic description which helps us understand how we think. But a straitjacket is a straitjacket, and no less restrictive based upon its reasons for being worn.

What can be said to this complaint on behalf of commonsense psychology? Three objections come to mind. First, it might be said that there is actually a very diverse range of propositional attitudes at play in commonsense psychology, and concentrating on just two is something of a straw man. However, this simply is not borne out by the philosophical literature, which concentrates almost exclusively on belief, with desires getting an occasional look-in. Besides which, our argument that body image and body schema cannot be integrated into commonsense psychology turned on the difficulties of ascribing propositional content, not any specific attitude.

[97] We can legislate that all nouns be replaced with either 'Tittifer' or 'Thribb' according to some policy or other – any will do. It would be a very poor system for dealing with the many and varied things in the world, missing many important distinctions, *but it would be a system*. Although a little more extreme, this is otherwise a fair analogy for commonsense psychology. To digress slightly, the limited mentalistic understanding of people with autistic spectrum disorders is widely explained as a deficit in commonsense psychological abilities, since autistic children of otherwise normal or high intelligence fail the false belief task (see Baron-Cohen). Based on my own experiences of working professionally with autistic children, as well as considerations detailed above about the false belief task, I think it much more likely that this developmental disorder interferes with sufferers' abilities to grasp social contexts in the first place, and they are forced to develop something like commonsense psychology itself: an awkward, stilted and detached set of rules for working out thoughts from behaviour and vice-versa which works after a fashion, but none too well at times.

Second, it could be objected that commonsense psychology as a theory works fine; its widespread use in philosophy and psychology proves that. But as we mentioned above, popularity is no reliable guide to quality. In the absence of any other evidence it might make for a weak inductive argument, but given the robust empirical evidence which undermines commonsense psychology this argument is worthless, no more than a snouts-in-the-trough fallacy. If everyone else is at it, how can it be wrong? Apparently this was a common refrain amongst City bankers prior to the recent recession, to say nothing of the scandal over MPs' expenses.

The third objection is more interesting: reading the very large literature on preferential looking and the core knowledge hypothesis, the theory itself is frequently expressed in terms of representations. How can it be used to undermine commonsense psychology, the descriptive basis for representational theories of mind, when it presupposes mental representations exist? In answering this, we need to return to the main purpose of this thesis – to argue that there are no distinct entities which bear the properties which RTMs claim mental representations possess. This doesn't mean no-one should be allowed to use the word 'representation' – just as long as they don't use it in anger! I have found no use of 'representation' in the literature discussed above which makes any clear commitment either to the existence of mental representations as distinct entities, not to any of their key features identified in chapter one. Spelke and Kinzler's use of 'representation' is typical: they write about 'core systems of object representation', 'core systems of agent representation' and so on. This might look like commitment to a representational theory of mind; after all, if a system represents objects, then presumably it must be representational. This is not the case. Even if it is the authors' intentions to couch the findings they discuss in terms of representations as a deliberate commitment to RTM (which is questionable), there is no reason to think the same findings cannot be cast in other terms. Nothing in the empirical studies themselves requires any commitment to the existence of mental particulars such as representations, much less ones which achieve their specific functions in virtue of to their internal structure. Casting the results in terms of representational systems is either an additional layer of interpretation which leaves the actual data unaltered, or else is innocuous in that it makes no commitment to any claims about the existence or nature of

mental representations. It might even be both. Most uses of 'representation' in psychology, and philosophy of mind for that matter, are ontologically innocuous as they make no requirement on the nature or even existence of mental representations, and are just a common by-word for thoughts which are related to some external object(s).

Commonsense psychology really is a poor description of how people think. If we want to get a really good descriptive account of how people think – including, but not limited to, how we understand each other – then we must turn to empirical studies *before* we decide the key features of our description. This was one of the principal methodological concerns we identified in chapters three and four, and its importance has been well illustrated here.

What we absolutely must not do is exactly what commonsense psychology is guilty of – impose from on high a theoretical framework which accords with our preconceptions about the subject matter, then look at the data through the tinted spectacles we have just put on. To have done so is, perhaps, a forgivable failing to some extent. After all, none of the empirical data surveyed here was available when the theory of commonsense psychology was being developed. And it is easy to imagine that other people tend to lead their lives the ways in which we do. If you deliberately assembled a group of people who – unusually – spend most of their time engaged in detached and highly theoretical armchair reasoning, just what sort of theory of mind would you *expect* them to produce?

5.7 Commonsense Psychology and Mental Representations

We have seen that commonsense psychology provides the descriptive characterisation of intentional thought upon which RTM is based. Through a combination of empirical and conceptual evidence we have seen that commonsense psychology is itself an inadequate characterisation of the way we actually think. Therefore, any analysis of thought which takes commonsense psychology as its starting point will necessarily be distorted and mistaken. This is the case for RTM, and as a result we must question the existence of mental representations themselves.

Recall that according to RTM every propositional attitude corresponds to a token mental representation (for Fodor, having a propositional attitude simply *is* tokening a representation). Our ability to think in the ways we do is attributed to the existence of mental representations, the horizontal relations which hold between them, and their internal structure which allows them to act as mental surrogates for the external objects they are isomorphic to. If thought does not actually consist in sequences of propositional attitudes, then mental representations are an analysis of something which does not exist, and so there is no reason to believe they exist either. Thought does not consist in sequences of propositional attitudes, so mental representations do not exist.

6. Thought without Mental Representations

In the previous chapter we saw that the descriptive basis for RTM – commonsense psychology – is false, and that as a result RTM's commitment to the existence of mental representations is mistaken. In this chapter we will see that it is possible to rescue the attractive aspects of Millikan and Fodor's theories – their semantics – whilst jettisoning any reference to, or requirement for, mental representations. The aim of this chapter is to show that Millikan and Fodor's commitment to mental representations is completely unnecessary.

We will achieve this by constructing alternative theories of mind which accommodate their semantics and their stated assumptions, yet have no place for mental representations. This will demonstrate that naturalistic semantics' supposed commitment to these representations is entirely illusory; it has no more substance than the Emperor's new clothes. Mental representations are redundant in a naturalistic theory of mind, and were never well-motivated in the first place.

In doing so we will commit the same fallacies as RTM: cavalierly imposing a highly artificial framework on the phenomenon it is supposed to analyse, dictating the nature and workings of the mind without due care and attention to a proper characterisation of just those things *prior* to their analysis, and being primarily motivated by highly theoretical philosophical concerns.

If the aim of this chapter were to build a sound theory of mind, it would be equally guilty of the methodological flaws we diagnosed in RTMs, as we will be imposing a descriptive ontology gerrymandered to meet our theoretical aims. However, our aim is quite different: it is to expose mental representations as unsupported and unmotivated by their own advocates. We adopt the same methodological approach as RTMs, the same assumptions, and the same semantics. But even given all this, there is still no requirement for mental representations. There is no work for them to do.

To demonstrate the redundancy of mental representations, even in the theories of their staunchest advocates, in essence we can cross out every instance of the word 'representation' without affecting either semantic theory at all.[98] In practice, of course, a little more needs to be done than that – but surprisingly little.

What needs to be done is to put in place a descriptive ontology which is as close as possible to that presupposed by RTMs, only minus mental representations. This is not entirely straightforward since, commonsense psychology aside, their ontology is largely vague and implicit – however, there are some clear points. Our ontology needs to be naturalistic, as Millikan and Fodor are quite explicit on this point. The exact meaning of naturalism can vary enormously – and as we saw earlier there is a lack of clarity in the literature on RTMs about its meaning in this context – but at minimum we will take the natural and non-natural to be continuous and allow no *super*natural objects or events. We will also restrict our ontology to what is physical – arguably this is entailed by naturalism, though I do think there is an argument to be had here – if only as a simplifying assumption. Apart from that we should have all the normal sorts of things – plants, planets, people, fundamental particles and so on. In discussing Fodor's theory, we will also address all six explicit assumptions he makes in his recent work *LOT2*.

But first Millikan. In short, if we accept vehicle externalism – which is both compatible with Millikan's views and attractive on independent grounds – then mental representations lose both motivation and purpose. For Fodorian semantics more needs to be done; an externalist variation on the identity theory of mind readily preserves all the main features of his semantics whilst leaving no role or requirement for mental representations. Neither of these variations contradict any explicit or implicit principle advocated by Millikan or Fodor, but they do rule out the existence of mental representations. We will pursue both these lines of thought below.

[98] It is my view that this generalises, *mutatis mutandis*, to all varieties of RTMs. In this thesis the argument is restricted to Fodor and Millikan.

6.1 Biosemantics without Mental Representations

In chapter one we saw that Millikan developed a sophisticated variation on Dretske's informational semantics, one which is essentially disjunctive in character. To summarise, when an object being thought about is physically present, the intentional content of a representational thought about that object is derived from the object itself. This process is, *mutatis mutandis*, essentially the same as Dretske's informational semantics, with a thought's content being provided by a reliable causal connection between the object being thought about and the person entertaining the thought. When there is no causal connection to the object being thought about – either by mistaking *X* for *Y*, or else when the object doesn't exist – Millikan departs from Dretske in her explanation of content fixation. In such cases the content of a representational thought is determined by the evolutionarily-determined biological functioning of a mental process. Together our various mental processes allow us to function normally in our environment (and generally not be mistaken about what is going on), and this normal (or 'proper') functioning is what explains the possibility of contentful thoughts without an appropriate objects to originate their content. Millikan (2008) uses the analogy of a coffee maker with the wrong ingredients. It won't make coffee, but it is still a coffee maker and it *is* functioning normally; it just lacks the right input to produce the right output.

One of the key logical features of any disjunctive theory is that by definition there is no need for a common factor in both cases. This can be seen clearly in contrasting conjunctive and disjunctive theories of perception. The conjunctive theory asserts that all perception – whether veridical or non-veridical – involves a perceptual experience. In cases of veridical perception the experience corresponds to whatever object or state of affairs is being perceived, whereas in cases of non-veridical perception the experience does not. In these latter cases, either there is a mismatch between the perceptual experience and what is actually there, or else there is nothing physically present at all. The parallel with misrepresentation should be clear.

In contrast, disjunctive theories of perception hold that there is no common kind of experience between these different cases of veridical and non-veridical perceptions. In fact, disjunctivists may be direct realists about perception, and so wish to hold that veridical perception is completely unmediated – an option not available to conjunctivists. Soteriou offers a detailed analysis of various types of disjunctivism including arguments for and against, and there is no need to repeat them here.

As a disjunctivist about representation, we might expect Millikan to likewise reject a common factor in cases of true (veridical) and mistaken (non-veridical) thoughts. There are some slight hints that she is sympathetic to this view, although to the best of my knowledge she does not endorse it as such. For example, she claims that teleological theories in general are not theories of content at all (2008, p.3 of online version). They need to be supplemented by a separate theory of what representations are, and how they function in non-mistaken cases, which Millikan supplies in the form of pushmi-pullyus and other intentional icons which inform and direct action in virtue of isomorphisms between them and whatever states of affairs they are about. However, there is no explicit rejection of any common factor between cases of true and false representation.

There is a strong pragmatic ground for an explicit rejection of a common factor for disjunctive teleological theories such of Millikan's, with the common factor being mental representations. Here is why. Mental representations are characterised as fulfilling their semantic function of explaining how thoughts can be 'about' particular objects or states of affairs in virtue of a structural isomorphism between the representation and the thing being represented. Though details may vary, we have seen this characterisation is true not only of Millikan's theory, but of RTMs in general. By definition, in cases of true representation there will always be an object present which the representation is structurally isomorphic to. But in that case we have to ask what role are the representations actually playing? It is a truism that everything is already structurally isomorphic to itself, so what is the value in positing a second isomorphic object when we have a perfectly good one already? For veridical cases at least, why can Millikan's semantics not operate solely on objects themselves?

The obvious answer to these questions is, of course, that the representation is supposed to be operated upon directly by mental processes, and physical objects or states of affairs aren't the sorts of things which are involved in mental processes. *Hence the need for a surrogate whose isomorphism to the object of thought preserves the relevant semantic features* (if you prefer to speak Fodor, you would say that external objects are not terms in mentalese, either simple or complex ones).

But this is a question-begging response, as it is by no means obviously true that physical objects or states of affairs *don't* participate in mental processes directly. It is the essence of vehicle externalism that they do exactly that.[99] Although Millikan is a content externalist (see e.g. her 2005 ch.3), I know of nowhere where she discusses vehicle externalism at all. I also know of no reason why her biosemantics should be incompatible with vehicle externalism . As we have seen in chapter one, extended-mind arguments make vehicle externalism quite compelling, assuming one is already a content externalist – and Millikan certainly is – so if the question were to be begged either way it would seem more logical to incline towards the view that 'external' objects and states of affairs can participate in mental processes.[100] Besides which, all RTMs take cases of veridical representation to be ultimately dependent upon the structure of the external objects themselves (remember, we are leaving aside all mistakes and reference to non-existent objects at the moment), so the move here is really only to focus on the original source rather than an imitation which reflects its salient points.

Given that objects of thought can participate in mental processes, a Millikanesque theory of mind has no need of mental representations in cases where the object of thought exists and is represented correctly. The objects themselves are quite capable of entering into semantic relations based upon their constituent structure, without the need for any isomorphic

[99] Incidentally, it is also part of the identity theory canvassed below that they do so.

[100] Here I am oversimplifying somewhat, as vehicle externalism itself does not conflate objects which participate in mental processes with the targets of whichever thoughts are operated on by those processes. We might well think that the two come apart in many cases. However, for present purposes we are only interested in a small subset of ways in which 'external' objects might participate in mental processes, that is being operated upon in virtue of semantic features which are derived from their composition. In such a case the two can be reasonably run together, at least as a simplifying assumption.

surrogates. We will now argue that Millikanesque biosemantics does not require representations for non-veridical cases either.

As mentioned above, Millikan has perceptively – and correctly – observed that teleological theories are not theories of content at all. They are theories about how a mental process or system can function when the normal process of mental content fixation goes wrong. This holds true regardless of whatever story is told about content fixation in the normal case – informational or not, representational or not. Millikan's teleology explains mistakes in terms of mental processes functioning in exactly the normal way, only in a different context which accounts for the mistaken nature of the thought – just like the coffee grinder with the wrong ingredients.

We can take apply this teleological strategy to our non-representational, externalist semantics. We have seen that in the normal case mental processes can plausibly operate directly on objects' structure rather than requiring a representational surrogate. In their normal, or 'proper', functioning mental processes don't require representations. Since the same processes operate in the same way in cases of non-veridical thought, there is absolutely no requirement for mental representations in these cases either. The mental processes still fulfil their 'proper' function – it's the inputs which are wrong. Ours is a coffee grinder just like Millikan's, only it's a little more economical.

To summarise: given that 'external' objects can participate in mental processes – and so there is no great metaphysical gulf between what is 'in' and 'out' when it comes to our minds and the rest of the world – there is no need for mental representations as surrogates of the objects we think about *when those objects are present*. Disjunctive teleological theories such as Millikan's biosemantics explain the existence of thoughts about non-present objects in terms of the normal or 'proper' functioning of (biological) mental processes. This teleological explanation can be applied equally to mental processes which don't operate upon representations to produce a new biosemantics which is functionally identical, except that there is no need or role for mental representations. The semantics are identical in both cases, the only difference is that by adopting vehicle externalism (which is

independently plausible) mental representations disappear. By Occam's razor, the simpler theory wins.

6.2 An Identity Theory

We can apply a variation on this technique to Fodor as well. As with Millikan the method is to adjust our descriptive characterisation of the mind slightly, but keep it naturalistically respectable. Since Fodor has recently stated six explicit presuppositions for his theory, we will match them point for point. The result is that commitment to or requirement for mental representations again disappears, leaving Fodor's semantics intact. This demonstrates that regardless of whether Fodor's presuppositions are true or false, he has no good grounds for believing mental representations exist, and his semantics function equally well without them.

This is our descriptive ontology: we will assume the mind is identical to the brain, at least some of the body, and parts of the more distal environment traditionally seen to be 'outside' of ourselves. The inspiration for this theory schema is traditional mind-brain identity theory (with the classic locus being Place 1956), the difference being that here the mind is identical to the brain *and* other aspects of the physical world as well. The relation would likewise be one of type-identity rather than token, meaning that minds and thoughts in general would be identical to physical states or processes as well as this holding true for some specific cases. This is not to say that thoughts are necessarily identical with their intentional contents, nor with their vehicles. Nor, for that matter their contents *and* their vehicles (though none of these are possibilities ruled out either). All it claims is that the mind, and individual thoughts entertained by it, are nothing over and above the brain, the body and some of the rest of the physical world.

Whether this schema could ultimately be developed into a plausible theory of mind is debatable. A key issue is specifying how the world is divided up into what is part of any given mind or thought and what isn't. This issue will have to be dealt with by a fully-fledged

theory along these lines; our purposes only require a schematic outline. Restrictions on how much of the body and the world are included seem to be inevitable: if literally *everything* was identical to the mind then there would only be one, which isn't a desirable result. So there must be some things which are not part of any given mind – and given what we believe about the size, age and composition of the universe, this probably includes the overwhelming majority of things. This probably extends to our bodies as well. It would be an odd result to say that clipping one's toenails resulting in a part of one's mind being lost.

Should the body have a role in the functioning of the mind? Our earlier discussion of the body image versus schema distinction is highly suggestive that it should, but is identity too strong a relation? It is true that brain damage can have mental repercussions, which is consistent with an identity theory like that being canvassed. It is probably the greatest attraction of traditional mind-body identity theories, and is preserved here. I see no good reason to stipulate that some other physical changes outside the central nervous system could not have a direct influence on the individual's mental processes; a dramatic change in one's body could conceivably lead to a change in the individual's way of thinking. This could, of course, be attributed to trauma or some other purely psychological reaction to the change, but in the context of our reductive theory there may very well be no need to stipulate trauma as a physical process on top of the physical process of bodily change. More likely, the two would be continuous and together sufficient to account for psychological changes.[101] If this turns out not to be the case, we can simply heighten the restrictions on what parts of the body enter into the identity relation with the mind. This is only a schema, after all.

The only real novelty of the identity theory here is a thorough-going rejection of the metaphysical significance of two traditional boundaries in the philosophy of mind: between body and environment, and between the central nervous system and the rest of the body. We briefly surveyed different forms of externalism in chapter one. They all have one feature

[101] Note I do not say necessary as well as sufficient: this would rule out at least some cases of psychosomatic illnesses where the trauma or other psychological reactions are present but the physical illness as a cause of them (rather than effect) was initially absent. Perhaps these could be explained purely in terms of the state of the central nervous system; more likely I think they will include features of the more distal environment. Without being overly Freudian, formative experiences could perhaps be included as relevant, hence part of the mental event.

in common, which is that a significant metaphysical boundary is maintained between what is 'inside' the body and what is 'outside' of it. Even vehicle externalists keep the distinction: 'extended mind' theories claim that cognition occurs not only inside the brain, but outside the body as well. We will reject the distinction entirely: there is no 'inside' or 'outside' at all. *This is not an argued claim*, rather it is a programmatic assumption. Of course the brain is vital to cognition in virtue of its physical composition and function, but from our perspective there is no principled difference in type between brains, bodies or anything else.
Other issues of detail also spring to mind, for example is there anything physical which is not part of a mind, either contingently or necessarily? Unfortunately there is not sufficient space to address this or many other issues a full theory would need to resolve. We will have to settle for canvassing this schematic version in enough detail to show it is fit for our purposes. Whilst on the subject of restrictions, however, a quick note: it may be wise not to restrict ourselves to identifying the mind with objects, events or processes which occur or exist entirely at the same time. This restriction would be absurd if it meant we could only think about what exists at the same time as us, even worse if we could only think about what exists at the same time as the thought. Any theory of cognition must allow cognitive relations to past and future objects and events, likewise ours here. The only difference is that our relation is a particularly strong one.

Whatever the status of this identity schema as a plausible theory of mind, our purposes here are much less ambitious. The theory need not be true[102]; all we need is to show that it can accommodate Fodorian semantics without commitment to mental representations, whilst also avoiding any egregious logical inconsistencies or contradictions.[103]

[102] In fact, some support for the schema could be drawn from Buddhist metaphysics, which has a long and highly respectable pedigree. That is not to say the theory here is Buddhist – or that Buddhists are identity theorists – rather that it is consonant with key Buddhist teachings. In particular *anatta*, the view that there is no unchanging self, and *dependent origination* which states nothing is ontologically independent of everything else. These teachings entail that there are no substances in the classical sense associated with Plato and Descartes amongst others, which is also a corollary of our identity theory.
[103] Latent logical difficulties could be permissible. After all, one way of expressing the overall argument of this thesis is that RTMs contain a latent oddity: they stipulate a gulf between individual people and the rest of the world, and concoct a class of curious entities to bridge it. Why not just drop both gulf and bridge? After all, sometimes philosophy *should* be therapy.

6.3 Fodorian Semantics without Representations

We will address those two tasks in turn. Recall that Fodor's semantic theory of asymmetric dependence is itself best viewed as being schematic. Our identity theory, which explicitly rejects mental representations as a distinctive class of entities, satisfies the schema. Cases of veridical representation are explained in terms of law-like, counterfactual-supporting causal connections, and mistakes are asymmetrically dependent upon these veridical cases. So mistakenly thinking a cow is a horse is dependent upon the normal case of correctly thinking of horses as horses. The converse does not hold, as thinking of horses as horses is in no way dependent upon thinking of anything else as horses.

The essence of Fodor's semantics for both veridical and non-veridical thought is that the former gain meaning in a law-like, counterfactually-supporting manner and the latter are dependent on the former but not vice-versa. Our identity theory satisfies both requirements:

What could better support a counterfactual than an identity statement? Either a given state of affairs (I use the term neutrally here) exists or it doesn't. When it does exist the appropriate content is *ipso facto* produced in normal cases, when it doesn't exist the appropriate content generally is not. Here existence and identity take the place of a causally reliable flow of information between thinker and object (using Dretske's clearer terminology, the same applies *mutatis mutandis* to Fodor's), and hence explain content fixation in veridical cases.

First objection: an identity relation could never produce an inaccurate mental content, so our schema entails that all content is veridical, so no thought would ever be mistaken. People do make mistakes, so the theory is false. This is not so – we explicitly ruled out that in general the content of a thought must be identical to its object, and there is no reason to think that this would be the case.

Second objection: there is no analogue in our theory for *un*reliable causal connections – ones which deliver information accurately only some of the time. This is likewise misguided. If our identity theory stipulated that existence of a subject and object was *all* that is required for cognition the objection would stick, but the theory absolutely does not. As presented it is incomplete; to yield a full account of cognition it would need to be supplemented with a theory of perception, one which tells us about the conditions under which something is seen, as well as under what aspect it is seen, and so made available to cognition. That much is far beyond the scope of this project, but I see no problem with assuming it is possible. One option is to advert to a set of reliable causal connections to make the necessary distinctions. In that case our new theory would be practically indistinguishable from traditional causal semantics as advocated by Fodor (and others such as Dretske and Millikan), *except that there is no requirement for mental representations*.

If Fodor's causal semantics for veridical thought is plausible, on the grounds that it employs a law-like relationship between the content of thoughts and their objects in normal cases, then our theory works just as well. Identity is a law-like relation *par excellence*.

6.3.1 Asymmetric Dependence

What about cases of non-veridical thought, and Fodor's asymmetric dependence schema? Identity is not asymmetric, and that's what is at the heart of Fodor's semantics, so surely our identity theory cannot satisfy his semantics? This problem only arises if we try to replace Fodor's dependency relation with identity, which is not what we are doing at all. What Fodor requires, and the identity theory accommodates, is that cases of e.g. cows causing horse-thoughts depend upon cases of horses causing horse-thoughts. To accommodate this we can appeal to relevant similarity; on our theory any given thought would derive its content holistically (as it were) from a broad state of affairs including the thinker and the broader environment (not just the object and/or target of thought). In two sufficiently similar circumstances we could reasonably expect the content of the thought produced to be the same, even if there were some salient differences (such as it actually being a cow

over in that field, not a horse).[104] Mistakes happen because cows look somewhat like horses, and in certain circumstances look a lot like horses. Our identity theory can explain cows and horses both producing horse-thoughts by sufficient relevant similarities between two circumstances – one where the object of cognition is a cow, one where it is a horse – yet explain why only one case is veridical appealing to identity between the object and the target of the thought.

I must admit that relevant similarity is a rather vague term, but we are using one theory schema to satisfy another here, so this can perhaps be excused. Besides which, the example requires that cows *are* relevantly similar to horses, enough so that one could reasonably be mistaken for the other. If this is unclear, re-read the example replacing 'cows' with 'windmills' and 'horses' with 'giants'. We are only making use of a resemblance which exists already. The fact that relevant similarity is a challenging concept to unpack does not alter its significance or use, either for our theory or for the everyday thought it is analysing. My argument is just that our identity relation can accommodate the asymmetric dependency of content required by Fodor, which it does.

So our identity theory schema satisfies the requirements of Fodor's semantic theory, without any commitment to or role for mental representations. Punkt, one might say. Incidentally, it also satisfies Millikan's biosemantics in the way described above since our theory entails vehicle externalism.

[104] After all, one thing that RTMs get right is that our cognitive grasp of the objects we think about is invariably partial and incomplete. Thoughts about any thing at all don't exhaust everything that thing is, but pick out salient aspects.

6.3.2 LoT: Compositionality, Referentialism and Atomism

But unlike Millikan's biosemantics, Fodor's asymmetric dependence schema is one aspect of a broader *metaphysical* picture of cognition, his Language of Thought (LoT) theory. If our identity theory is dramatically at odds with the main tenets of LoT, there might be some justification to arguing that it begs the question against Fodorian semantics as an *overall picture* of cognition, even if it does satisfy his asymmetric dependence requirement. Luckily our theory is fully compatible with all six assumptions Fodor himself identifies as required for his theory to work (2008, pp. 198-200), so the question begging charge fails. If the assumptions are true, our identity theory accommodates them just as well as Fodor does. If any are false, our theory fares no worse than his. Either way, there remains no need or role for mental representations. We will address the assumptions in turn.

1. *The semantics of thought is prior to the semantics of language*.

In favour of this assumption we can see that from Spelke's core knowledge hypothesis that at least some of the semantics of thought definitely are developmentally prior to language development (certainly to the ability to deploy language through speech). While I have no intention of committing myself fully and unreservedly to the priority of thought over language, it should be clear enough that our identity theory is at least compatible with its assumed truth. It is actually a methodological assumption which Fodor is making here – if the semantics of thought is not prior to that of language, then his study of semantics needs to take a very different course to the one it has. In particular, the meaning of expressions (whether thoughts or utterances) is likely to be determined in a manner which is to some extent relational and non-individualistic, or 'pragmatic' to use Fodor's grossly broad umbrella term. My own opinion is that his dichotomy of his own 'Cartesian realist' versus 'pragmatist' theories distorts and oversimplifies a complex spectrum of views, but for the sake of argument we can accept it and side with Fodor. Nothing in our identity theory commits us either way.

2. *Reference is compositional*.

In other words the reference of a complex expression is composed from both the referents of its simple parts, and their manner of arrangement.[105] Fodor's view is that this is absolutely central to his theory: 'capturing the compositionality of thought is what RTM most urgently requires' (2008, p.17). By his own reckoning Fodor is part of an intellectual tradition dating back to Aristotle through Descartes (2008, pp. 5 & 9 respectively) and also Hume, with his RTM being 'more or less interchangeable' with Hume's Theory of Ideas (2003, p.8 fn.2). The major difference (nativism aside) being that Hume's associationism is semantically transparent – Fodor claims that it cannot differentiate between the content of A-associated-with-B and the content of A associated with the content of B. So associationism cannot distinguish between the single complex concept MRJAMES BITES and the sequence of concepts MRJAMES and BITES (Fodor 2003, ch.4). However, a compositional semantics can recognise both the content of simple ideas and their manner of arrangement in determining the content of a complex idea, and hence can distinguish each case clearly. Compositionality can be seen as effectively adding brackets to the expressions above: ((MRJAMES) BITES) versus (MRJAMES) BITES.[106]

Fodor's RTM explains how compositionality is implemented via the view that the interactions between representations are computational (his CTM). Certainly classical computation is compositional, hence so is Fodor's RTM. Our identity theory cannot follow suit since it denies there are any representations to enter into computational relationships in the first place. However, our theory does require compositionality on independent grounds. It holds that what is required for successful reference of a complex expression is the existence of a referrer plus some objects in the 'external' environment to which the referrer is related in some appropriate way (as yet unspecified, although we'll return to this briefly in considering Fodor's fifth presupposition below). In veridical cases, the object(s) in question will be the one(s) being referred to, if not then it must be something else. It seems utterly impossible to me that in this scenario reference could be anything but

[105] Contra classical mereology, Fodor is quite explicit in taking composition to be structured, and makes this point repeatedly in his 2003 and 2008.
[106] Here I have added a single set of brackets for each complex idea, with none for simple ones. MRJAMES is treated as a complex concept.

compositional. After all, reference is determined by the identity conditions of multiple objects themselves plus some plausible criterion of arrangement which distinguishes which are being targeted and which are not. As long as the identity conditions of complex objects are in themselves compositional – and I take it as beyond question that they are – reference will be as well.[107] So our identity theory requires Fodor's second assumption to be true.

3. Referentialism is true.

That is to say the content of a thought is entirely determined by what it refers to (for simple concepts, with complex concepts composed from simples as above). There is no problem with Fregean senses, because they don't exist. In some cases this is relatively easily explained: many coextensive concepts can be differentiated by their constituent structure, e.g. water is simple, whereas H2O is complex (Fodor 2008, p.66). As such they plausibly have different possession conditions – the concept of water may be acquired perceptually, whereas the concept H2O requires prior knowledge of concepts for hydrogen and water, plus some basic grasp of chemical composition. Harder to account for are basic concepts which are coextensive, e.g. Cicero = Tully. Since both may plausibly be acquired perceptually the same strategy will not work. Fodor's solution is that the causal powers of the two expressions are determined by their syntax in mentalese (ibid., p.68); after all, if the semantics of thought is prior to that of language then mentalese must have at least as detailed and sophisticated a syntax as English, if not more so.

Let's assume that both of these arguments are sound; in brief, that Frege-style cases can be explained away by appealing either to the structure of the concepts themselves, or of the way they are instantiated in the brain. Can we make a comparable move using our identity

[107] Maybe this claim needs a little more support. Imagine a pile of bricks: its identity conditions plausibly depend on those of the bricks it is composed of plus their manner of arrangement. Quite *how* we cash this out is a trickier question, as our consideration of mereology has made all too clear, but to say that the pile is not essentially made out of the bricks plus their arrangement – *in some sense* – is quite bizarre. Unless, of course, we don't believe in bricks – but we would say that it is made out of fundamental particles instead, in which case the same point applies *mutatis mutandis*. Classical mereologists also would disagree with the claim that physical arrangement is a relevant criterion, but in that case we could use a temporal restriction on the bricks/fundamental particles to much the same effect. Perhaps there are examples of complex objects whose identity conditions are not compositional, but I am currently at a loss to think of any. An object whose parts are in a state of quantum entanglement might just fit the bill, but even this exotic situation is open to interpretation.

theory? Again, the answer is definitely yes. For complex concepts we can simply adopt Fodor's appeal to constituent structure *tout court*. For simple concepts we cannot appeal to mentalese syntax, because without mental representations we lack both motivation and a likely medium for any such language of thought. However, our identity theory schema commits us to viewing any mental activity whatsoever – even entertaining so-called 'simple' concepts – as being constituted by a complex of objects. Imagine someone sees his friend John at a distance and correctly identifies him as such, and for the sake of argument thinks 'There's Tom' to himself. According to our identity theory schema this requires at minimum that the person should exist, likewise Tom (for it to be a correct identification), plus satisfaction of some referential criteria, probably perceptual in nature. The thought 'There's Tom' is nothing more, or less, than this.[108] This (or a fleshed-out version of this outline) will necessarily be a complex object.

Now for coextensive terms. When reading a book published under Tom's pen name, the same person's thoughts about the author will be constituted by an overlapping yet quite different complex of objects, most likely including the book and possibly John again. Exactly what is 'in' will no doubt be affected by whether or not the reader is aware of the author's true identity. Nevertheless, on our theory there are clear structural similarities and differences between thinking about a perceived Tom and Tom via a different name, which could be used as the basis for distinguishing between the coextensive uses of different proper names to refer to refer to the same person.

Because of this we can modify Fodor's strategy for dealing with simple concepts – instead of appealing to structural differences in complex internal objects (the syntax of mentalese terms), we can appeal to structural differences in the complex objects which transcend the internal/external dichotomy and which (we claim) constitute thought. Apparently simple concepts such as Cicero or Tully are actually complex insofar as they are constituted by complex objects. Whether this is actually true is moot; our identity theory can assume referentialism just as well as Fodor can, and on essentially the same grounds.

[108] This is, of course, only the skeleton of what would need to be in place in a relatively developed version of the theory. I suspect that a significant number of other objects would need to be included as well, ranging over a significant period of time. What they might be is beyond the scope of our discussion here.

4. *The vehicles of reference are exhaustively singular terms and predicates.*

As this conceptual atomism is a corollary of assuming both referentialism and compositionality (Fodor agrees here, see his 2008 p.20), we need say little about it. Since our identity theory has been seen to be able to assume both principles, it can readily accept this as well.

5. *'Some sort' of causal theory of reference is true.*

We have already seen that without a theory of reference, our identity theory of mind is incomplete. In our semantics existence is a surrogate for information, and suffers from the same problem of ubiquity. Information is everywhere, as are things (meaning whatever exists), so some restriction is needed to explain how reference is a more selective affair. Neither the presence of information nor existence of objects can be sufficient for reference, as the result would be an undifferentiated panpsychic mess.

Fodor favours a hybrid causal theory, according to which the reference of a singular present tense expression is 'fixed' by an act of ostensive definition, with subsequent uses 'borrowing' their reference from the initial case. Some mentalese expressions however may be fixed by definite descriptions, in particular those which are not expressing present-tense perceptions (2008, p.200). He devotes most of the final chapter in his *LOT 2* to defending this view. We don't need to evaluate the arguments involved, however, as once again our identity theory can accommodate this assumption. There is nothing to prevent us doing so. It is worth noting that Fodor's commitment to assumptions *2*, *3* & *4* seems to motivate his adoption of a causal theory. After all, if meaning is exhausted by the reference of singular expressions (plus their composites for complex ones) then he will need to adopt a theory of reference which is concerned with the fixing and continued use of singular terms individually (*3* & *4*), and in a manner which is unmediated by descriptions or anything else (*3*). The descriptive element is designed to take care of standard objections. Likewise, our identity theory will have the same motivation. My own view is that while assumptions *2*, *3* & *4* do not logically necessitate a hybrid theory of reference, they are quite consistent with one.

6. *The crux of the problem of naturalising reference is to provide a theory of perceptual reference.*

Again, this is motivated by other assumptions Fodor makes. Once the reference of singular terms is explained, all else will follow (by *2*, *3* & *4*), and given a hybrid causal theory of reference (*5*) the paradigm case to be explained is the initial fixation of the reference of singular expressions. It seems likely that in the main this will be achieved perceptually, by seeing something new and referring to it (especially since the reference of non-present expressions may be fixed by description).

Wittgenstein was right to argue that language cannot be built upon ostensive definition alone, and much of the empirical evidence considered in the previous chapter supports the view that there is significant understanding of the world at a pre-linguistic age. Nevertheless, some period of content fixation for singular expressions is required to ground Fodor's enterprise.[109] It may be that just such a period is a developmental stage in infancy, though there seems relatively little evidence to support this as yet. But whether or not there is a discrete developmental period of ostensive definition, and regardless of what cognitive abilities it presupposes, it is clear that a naturalistic theory which assumes 1 – 5 is dependent upon a satisfactory theory of perceptual reference. Fodor does not have a complete theory to hand, but does devote one chapter of *LOT 2* to a relevant issue: preconceptual representation.

In parallel to *1* above, Fodor explicitly assumes that representational thought is prior to perception – but 'nothing precludes the possibility that some of the *representing* that goes on in seeing/thinking is nonconceptual' (2008, p.169). To accommodate this possibility, Fodor distinguishes between 'iconic' (nonconceptual) and 'discursive' (conceptual' representations. The distinction between the two lies 'turns on difference between the ways that they achieve their compositionality' (*ibid.*, p.171). Unfortunately, this difference turns out to be one of degree rather than type, which greatly undermines Fodor's distinction.

[109] Here a latent similarity with Dretske comes to the fore.

Discursive (conceptual) representations undergo 'canonical decomposition'. What this means is that some of their parts are well-formed, i.e. are concepts, and others aren't. Take Fodor's example of 'John loves Mary'. It has canonical parts of 'John' 'loves' and 'Mary – each being concepts themselves – but other parts are not concepts, e.g. 'John loves', 'loves Mary' and 'John … Mary'. Examples of non-canonical parts include any combination of letters which don't constitute a concept themselves.

By contrast iconic (nonconceptual) representations can be divided up any way you like, and their parts are still iconic representations. Fodor uses the example of a picture, assimilating the majority of cases of iconic representation to this type. Cut a picture of something in half and the two pieces are still pictures (only now they are pictures of parts of the original something). Do it again, and again, and you still have pictures. Pictures, and iconic representations in general, have no canonical decomposition (2008, p.173).

In this way Fodor gives a neat characterisation of the difference between conceptual and nonconceptual thoughts – it's a difference in structure, i.e. in parts. Unfortunately every material thing has a canonical decomposition on his use of the term, rendering the distinction a blunt instrument at best.

Cut a picture of something up enough times and you don't have pictures of parts of that something, eventually you have a load of tiny bits of paper which are pictures of nothing. Keep going and you would end up with molecules, then atoms, eventually fundamental particles (whatever they might be). Regardless of whether we could keep going forever, and we probably can't, there is simply no way that the atoms (or whatever) which make up a photograph (or other picture) are themselves pictures of parts of the scene in the original photograph. Pictures, and other iconic representations in general, do have a canonical decomposition - it is reached at the point at which further decomposition renders their parts no longer pictures.

First objection: someone might argue that actually the atoms (molecules, fundamental particles, tiny pieces of paper etc) which make up a photograph do in fact bear a striking resemblance to the atoms (or whatever) which made up the scene being photographed.

They are iconic representations after all! Our reluctance to call them pictures only comes from associating the term with larger items like photographs, portraits etc.

This cannot be right. First of all, resemblance might be necessary for picturing, but it is not sufficient. Many things resemble something else without being in any way a picture of them. Resemblance is a symmetric relation whilst picturing is not. Worse, the claim is only true to the extent that *all* atoms (or whatever) resemble each other. If one is a picture of another, they are all pictures of each other, and that simply isn't what pictures *are*.

Second objection: the claim that pictures have a canonical decomposition is a trivial point about the nature of paper, or canvas, an LED screen or whatever a physical picture is displayed upon. Thoughts are not 'painted' on any such medium.

But they must exist in a medium of some kind (and general consensus has it that it is a physical medium within the brain). While it may well be a metaphysical possibility that some medium exists which can be sub-divided endlessly, assuming physicalism (as Fodor does, see e.g. his 2008 p.196) it is almost certainly false. I don't know how small a picture, or any iconic representation, needs to be cut up before its parts cease to picture (or represent) anything, but I do know this: *make enough divisions and it will happen.* Everything which exists in a physical medium has a canonical decomposition.

Fodor's distinction between discursive and iconic representations in terms of canonical decompositions is one of degree whereas the distinction between conceptual and non-conceptual thought is one of type, making the two fundamentally mismatched. It is possible that in some cases a difference in degree could underlie a difference in type, but this is not one of them. Conceptual and non-conceptual concepts are broadly similar insofar as they are both concepts, but on closer examination are quite distinct types. Iconic and discursive representations are superficially very different, but on closer examination turn out to be fundamentally the same in their structure. This makes the latter fundamentally unsuitable as an explanation of the difference between the former.

Given this fundamental flaw, Fodor's views on perceptual reference need some revision. A full and adequate theory is a long way off. It is nevertheless clear that the ultimate acceptability of his theory relies, at least in part, upon having a satisfactory theory available. So, too, will ours for exactly the same reasons. In its absence we are in no worse a position than Fodor is.

6.4 Thought without Mental Representations, Again

We have seen that our identity theory can satisfy not only Fodor's semantics in the guise of his asymmetric dependency schema, but also satisfies all of the assumptions he states are necessary to support his theory. Despite this, it has neither a requirement nor a role for mental representations. So assuming that our schematic theory suffers no egregious logical errors or inconsistencies, a question we will address below, advocates of RTM are left in an awkward position.

On the one hand we saw in the previous chapter that there are very good reasons to doubt that representational theories of mind are true, insofar as they presuppose an utterly bogus description of the phenomenon they seek to analyse, then cavalierly formalise it anyway. They are simply methodologically unsound, and this casts grave doubts on any claims they make. To accept this argument is to reject RTM, and any motivation to believe in mental representations as a class of entities vanishes.

On the other hand, even if we accept the abstracted, top-down methodology of RTM, we have seen further difficulties arise. In this chapter we have denied no assumption or tenet of either Millikan or Fodor's semantic theories, yet by making some minor adjustments to their descriptive ontology (vehicle externalism and an identity theory of mind), we lose any requirement or role for mental representations whilst preserving all other features of their theories.[110] What this shows is that

[110] But surely there is a glaring error here: our identity theory, while satisfying all of Fodor's assumptions, does not support a language of thought. Mentalese would seem ill-motivated under such a theory; the whole point of a language of thought is that it is fully internalised, and our theory denies the existence of any suitable

(1) for all its pretensions to clarity and rigour, RTM is a confused mess which distorts the very subject it proposes to analyse, and
(2) despite this the more attractive aspects of RTM can be retain *salva veritate*, whilst abandoning the central claim that mental representations exist.

Commonsense psychology is an inadequate descriptive characterisation of intentional thought, and this renders RTM invalid as an analysis of intentionality. But even if you are a naturalist, even if you are a physicalist, even if you think commonsense psychology is adequate after all, and even if you believe in one or another version of causal semantics – you still lack any need or motivation for mental representations.

The only way to avoid this conclusion is to reject our identity theory schema, or for Millikanesque biosemantics to reject vehicle externalism. We have already argued in chapter one that vehicle externalism is independently plausible. Before moving on we are still left with the question of whether our identity theory schema is at least plausible (if not necessarily *true*). This requires our theory should not contain some egregious logical error or inconsistency. It should, at least, be able to deal with the typical objections levelled against traditional mind-brain identity theories. Perhaps not deal with them to the satisfaction of all, but at least to the degree that a fairly plausible reply is on the cards. We will briefly consider the two principal objections, and argue there is an at least reasonably promising response to each.

---

internal medium. Surely this means it fails to preserve the single most significant aspect of Fodor's work? I don't think so. My view is that there is an identifiable development in Fodor's work of a metaphysic of mind which revolves around the composition of concepts entirely independently of what medium they are instantiated in. This can be seen in, for example, his powerful arguments for using structure in support of semantic referentialism. The hypothesis of Mentalese becomes increasingly peripheral to this overall work, and may yet be abandoned entirely. Fodor abandoned his semantic internalism, and I am of the opinion that he should abandon his vehicle internalism likewise. That this is possible is demonstrated by our identity theory satisfying all of Fodor's metaphysical assumptions, yet neither supporting nor requiring a language of thought.

6.5 Standard Objections to Identity Theories

6.5.1 Multiple Realisability

This argument against mind-brain identity theories was raised by Putnam (1967), who argued that it is plausible that mental states are multiply realisable, meaning that two life-forms could be in the same mental state without being in exactly the same physical state. This is widely held to be intuitively true (accommodating multiple realisability is widely seen as one of the major strengths of functionalism, which only accepts token identity), yet is apparently inconsistent with mind-brain type identity theories. There are numerous detailed and sophisticated replies to Putnam, which we don't need to repeat here (for summaries see e.g. Bickle, Smart 2008).

We will consider one line of response: it may be that mental states are not multiply realisable (see e.g. Zangwill), in which case Putnam's argument fails. This is not a very widely held view, but it can be used to argue that while different physical states may constitute the same mental state, this is not a genuine case of multiple realisability as the differences between them are irrelevant to their function *as that mental state*. Shapiro uses the analogy of a corkscrew: 'steel and aluminum are *not* different realizations of a waiter's corkscrew because, relative to the properties that make them suitable for removing corks, they are identical' (2000, p. 644). Likewise, a type identity theory can accommodate physical variation without mental variation – either the variation does not affect their mental function, in which case there is no multiple realisation as there is no *relevant* difference, or else the variation does affect their function, in which case they are different kinds – and so there is still no multiple realisation. Our type theory gives extra wriggle room as the physical changes in question are not necessarily restricted to the brain.

Shapiro's argument is sound. Multiple realisability may be something of a sacred cow in contemporary philosophy of mind, but from that it does not follow that it doesn't need to be argued for. Popularity is not necessarily an accurate guide to truth (remember commonsense psychology on this point). There may well be limitations on the multiple

realisability of the mental – whether partial or total – and it is far from clear that the debate over this issue has been played out yet. Our identity theory can make use of Shapiro's analogy and so fares no worse than traditional type identity theories on the score of multiple realisability, which is to say that at worst the issue is as yet undecided.

6.5.2 Phenomenal Properties and Qualia

A reductive physical description of a mental state or a thought can seem to leave out an important element: what it is like for the person who actually entertains it. The experience of seeing a beautiful sunset seems to be utterly left out by a description of various states of the central nervous system which, according to mind-brain identity theories, are all that the experience actually is. Typically referred to as qualia, these phenomenal aspects of our mental lives are notoriously resistant to physical reduction, and feature heavily in arguments against reductive physicalism.

Smart's (1959) response was to claim that the properties of experience are 'topic-neutral', borrowing the idea from Ryle's characterisation of words such as 'if', 'and' or 'not' which express structural relations between topic-specific terms. Smart's use is somewhat different in that it is restricted to neutrality between physicalism and dualism: he claims that the language used to describe experiences is neutral between these metaphysical theories. Thus 'I see a yellowish-orange after-image' means something like '*There is something going on which is like what is going on when* I have my eyes open, am awake, and there is an orange illuminated in good light in front of me' (Smart 1959). To explain how there can be *something going on* for me regardless of what may or may not be going on outside me, the infamous 'sense datum' – a mind-dependent object of perception which bears the properties which perceptually appear to us – is invoked.

Smart's topic-neutral response, and similar variations, suffer from a degree of vagueness in the use of 'something going on', upon which their main plausibility lies, a point which has been strongly pressed by Chalmers (1996, p.360).[111] Our earlier discussion of formal ontology allows a slightly different interpretation of Smart which produces an interesting response. Recall that we characterised formal properties in two different ways. One is that they can be applied cross-categorially, the second that they are topic-neutral. Examples of these properties such as parthood, representation, dependence and identity all express what might be described as structural relations between members of various different categories (identity is, of course, the formal relation *par excellence*: it applies to everything). Could the properties of experiences be the same? On a traditional mind-brain identity theory this seems a hopeless proposal. How could structural features of the brain possibly determine the phenomenal properties of our experiences? For one thing, this would completely ignore the issue of whether there is actually something 'out there' as a target for the experience in question. As with thought in general, it is an essential feature of experiences that they may be veridical or non-veridical, and this difference lies solely in their relation with what they are experiences of. Solipsism aside, there is no way the brain alone could satisfy the distinction.

However, our broader identity theory allows phenomenal properties to be structural features of not only the brain but also of 'external' objects. Thus, an experience of something entirely absent can be explained in terms of structural relations between the brain and other things which are there which suitably mimic the relations that would be in place were the experience veridical. Different experiences of one and the same object could be explained by structural differences in the two mental systems which both incorporate the same object, but different brains and/or other things besides. These variations could as easily be between the same person at different times or different people at the same time. Or, indeed, different people at different times. Not forgetting as well that the object(s) in question could be encountered in different ways; contemplating a vase is quite different

[111] Also, from Smart's topic-neutrality there is a lack of empirical evidence to favour type-identity over dualism, or vice-versa. Smart's own view, as I understand it, is that Occam's razor would favour his simpler explanation.

from being hit over the head with it. There is a wealth of ways in which differences in an experience could be attributed to structural differences in the physical world.

I would suggest this is a promising line of response, at least sufficiently so to argue that the existence of phenomenal properties is not a knock-down objection to the theory we are canvassing here. Like for traditional mind-brain theories, experiences are not *things*, not part of the 'furniture of the world'. Rather than advert to sense data to cash out an idea of *something going on*, however, we can take phenomenal properties to be not a question of what there is, nor indeed of how we see things, but rather of *how what there is is arranged*. We can illustrate the difference here by considering the Mary's room thought experiment:

> Mary is a brilliant scientist who is, for whatever reason, forced to investigate the world from a black and white room via a black and white television monitor. She specializes in the neurophysiology of vision and acquires, let us suppose, all the physical information there is to obtain about what goes on when we see ripe tomatoes, or the sky, and use terms like 'red', 'blue', and so on. She discovers, for example, just which wavelength combinations from the sky stimulate the retina, and exactly how this produces via the central nervous system the contraction of the vocal cords and expulsion of air from the lungs that results in the uttering of the sentence 'The sky is blue'. [...] What will happen when Mary is released from her black and white room or is given a color television monitor? Will she learn anything or not?
>
> Jackson 1982, p.130

The thought experiment can be interpreted either epistemologically – as being about Mary's knowledge – or ontologically about the existence of non-physical facts. Since our interest is in ontology we will take the second, stronger interpretation (which is probably what Jackson had in mind anyway). Accordingly, the thought experiment is designed to demonstrate there are non-physical properties, hence that physicalism is false, by inviting the answer that Mary will indeed learn something new. It is possible of course to give a negative reply as Dennett does (1991, p.398), since if Mary knew *everything* about colour she would know what to expect when she saw red. But Jackson's formulation above does not specify she knows everything, just that she knows all the physical facts. To assume that this is exhaustive is to beg the question in favour of physicalism.

There are many other objections, most notably that the knowledge Mary acquires is non-propositional, and better construed as either an ability (a view to which Jackson himself ended up subscribing, see his 2003) or else as some form of direct acquaintance. Arguments for and against both these views are well presented in Nida-Rümelin, alongside others, so I see no need to rehearse them further here. Our view is compatible with either hypothesis, though perhaps naturally closer to the acquaintance view. If phenomenal properties really are just structural features of collections of physical objects, then what happens when Mary leaves her room is very simple. By this act the composition of Mary herself changes, resulting in some different structural relations between her various parts. This might equate to her experiencing a new colour, or it might not.[112] Either way, our theory here has the makings of its own response to the thought experiment, or else can adopt one already present in the literature.

Inadequate and brief as they are, these responses to the two standard objections to identity theories are sufficient to show that our identity theory schema at minimum has enough plausibility not to be dismissed out of hand. The schema is fit to support the arguments of this chapter.

[112] Interestingly, Ramachandran & Hubbard use empirical studies of a colour blind synaesthete to suggest that Mary would have a form of 'blindsight' for colour whereby she lacks any experience of red versus gray, but can nevertheless sort objects according to the colours she has no subjective acquaintance with. This is still compatible with our structural view of phenomenal properties, as the structural changes which occur in Mary's physical composition may not include any new phenomenal properties, although there would be non-phenomenal ones. This would of course be unusual compared to most people's reactions, but if this seems to be an *ad hoc* response, remember that Ramachandran & Hubbard's analogy is with someone who confuses numbers and colours despite being colour blind. Unusual responses to perceptual stimuli are par for the course.

6.6 Summary

Millikanesque biosemantics has no requirement or role for mental representations, as long as we accept vehicle externalism. This principle is fully compatible with Millikan's views and is plausible on independent grounds. Fodor's asymmetric dependency schema can also be satisfied without any requirement or role for mental representations, by adopting an identity theory of mind which satisfies all of his metaphysical assumptions and can suitably address standard objections. Causal semantics should be properly viewed as separate from the RTMs with which they are commonly associated, and so provide no support for the view that mental representations exist. Not even Fodor and Millikan – representationists *par excellence* – have good reason to believe that mental representations exist. This is, of course, because they do not.

7. Conclusion

Our overall aim in this thesis has been to contribute to the study of the ontology of mind by pruning away one unnecessary and misleading entity – the mental representation. Though often used as a convenient byword in philosophy of mind, their existence as a class of mental particulars with unique properties is posited by a school of thought exemplified by Millikan and Fodor, called the Representational Theory of Mind. Mental representations are held to be mental entities which act as surrogates for worldly objects being thought about, and supposedly do so in virtue of isomorphisms between their internal structures and the structures of the worldly objects being thought about. This is designed to solve the problem of intentionality, by explaining how thoughts can be 'about' objects.

Representational theories admit considerable variation in their accounts of how the content of mental representations is fixed – in particular, Millikan emphasises the importance of the use to which representations are put while Fodor makes use of the syntactic structure of the representations themselves. The theories also explain how thoughts can be mistaken in a range of ingenious ways. Common to all variations is a commitment to the characteristically ontological claim that mental representations *exist*.

Representational theories of mind need to be assessed from an ontological point of view, both because they have ontological commitments and because they are themselves formal ontological theories, with the commonsense psychological view of though consisting in sequences of propositional attitudes providing their descriptive basis. Ontological methods and concerns are characteristically quite different from the semantic approach which characterises previous debates over mental representation. In order to illustrate what the methods and concerns of ontology should be (as which view is correct is a controversial issue), we undertook an extended investigation of the formal relation of parthood. Parthood is regarded as one of, if not the, most significant formal relations. It is intimately related to representation as the concept of internal structure requires that of parts.

We have seen that the dominant contemporary view of parthood is classical mereology, a formalised logical theory closely related to set theory. Various objections notwithstanding, we have seen that its axioms of reflexivity, antisymmetry, transitivity, extensionality and unrestricted composition are all defensible, at least when combined with some background assumptions. Whether classical mereology can be considered true depends in turn upon whether these assumptions – in particular perdurantism – are true. The answer lies outside this thesis, though I am personally rather sceptical. The supposed uniqueness of mereology as the *only* parthood relation is false.

Classical mereology fails to adequately address any of the problem cases of composition we set as a challenge for an adequate analysis of parthood. By supplementing mereology with elements of topology, the qualitative study of space, we saw that one of our problem cases concerning undetached parts could be adequately addressed. Further supplementing this mereotopological framework with concepts concerning function and causal stability is plausibly sufficient to address the other two.

In this way a heavily supplemented mereology can be seen to provide an adequate analysis of parthood, though only when combined with a popular but controversial characterisation of material objects as perduring rather than enduring. Fine's rival theory of embodiment adequately addresses all of our problem cases, but only if we assume that material objects endure. Mereology and the theory of embodiment were identified as contemporary varieties of two traditional doctrines, those of actual parts and potential parts.

Our analysis of these theories of parthood provides both an illustration of the distinction between descriptive and formal ontologies, and a vindication of its methodological value. Our evaluation of mereology would be difficult if not impossible to formulate without the distinction, and it has proved invaluable in clarifying areas of confusion – particularly over mereological extensionality. Our first methodological moral is that ontological investigation should both recognise and make use of the descriptive versus formal distinction.

We have also argued in favour of the so-called 'synthetic' approach to ontology which begins with empirical data (whether scientific or drawn from personal experience) concerning the domain or concepts of interest, and seeks to formulate theories based upon that data as evidence. The value of this approach has been repeatedly demonstrated throughout our analysis of mereology, and we have made effective use of examples from missing door handles and meandering rivers to infant development and molecular isomerism. These examples have not been mere illustration, but have guided and shaped our arguments and analysis. The necessity of drawing information from empirical data first, and only then formulating a philosophical theory as analysis of this data, is the second moral we drew from our discussion of mereology.

We applied our two methodological morals to the Representation Theory of Mind by questioning whether its descriptive basis – commonsense psychology – constitutes a sound descriptive account of everyday intentional thought. We made considerable use of empirical data – in particular studies of infant development, mirror neurons and neural impairment – to guide and inform our analysis. We found that commonsense psychology is a crude and restrictive analysis of thought, which requires a considerably richer conceptual repertoire to characterise effectively. We identified two concepts – body image and body schema – which should be included in a sound descriptive characterisation of how we think. The failure of commonsense psychology as a descriptive characterisation of intentional thought renders any formal analysis based upon it invalid, and hence the Representational Theory of Mind is false. Given its faulty foundations it necessarily distorts the very phenomenon of intentionality it seeks to explain. This provides a compelling reason to reject the theory's central claim that mental representations exist.

Finally, we argued that even the staunchest advocates of the Representational Theory of Mind lack adequate motivation to endorse it. The principal attraction of Millikan and Fodor's variations of the theory is their semantics – the way in which they explain how thoughts derive their meanings. We have shown that their respective semantic theories can be liberated from any commitment to – or even any role for – mental representations. Doing so requires no alterations to the semantic theories themselves, only some minor and independently plausible changes to the descriptive ontology informing the semantics. These

changes are fully compatible with the naturalism Millikan and Fodor endorse, as well as all of their other stated assumptions. Mental representations are shown to contribute nothing to causal semantics, which function exactly the same without them. Even if we were to reject every other argument of this thesis, there is still no compelling reason to believe mental representations exist, and so by Ockham's razor they should be abandoned.

If there were no other way of understanding the functioning of the mind, we should perhaps shrug and accept the Representational Theory of Mind despite all its faults, but alternatives are legion. I do not know how to accurately characterise the myriad workings of the mind, but I have a good idea where to start.

A possible line of future research I intend to follow is to further pursue the core knowledge hypothesis and its implications for the philosophy of mind, in particular understanding how our ability to refer develops in early infancy and what its development can tell us about the mature ability. By far the best piece of theoretical writing on this subject is Quine's *The Roots of Reference*; however while few of its claims have been refuted, it is substantially out of date given recent empirical data such as that produced by habituation studies, infant looming and mirror neurons. Studies in these areas, and many others, provide a wealth of evidence to correct, develop and in some cases supersede Quine's account of the origins of reference. Furthermore, the core knowledge hypothesis indicates that doing so may usefully inform our understanding of full-blown reference. Recent empirical data has the potential to transform philosophical analysis of reference, and related issues in the philosophy of mind, logic and metaphysics, but the data has not been thoroughly analysed yet. I intend to undertake this analysis.

As we mentioned earlier, it would also be highly worthwhile to analyse the concept of internal structure employed by RTMs using the formal ontological tools we have explored. It was at one point my intention to do so here, but unfortunately at the raw material to do so is currently lacking. If RTM should ever be developed to include a substantive account of *how* representations are structured, that account should properly be analysed using the best accounts of parthood and structure available, which we have seen to be either a

supplemented mereology or Fine's theory of embodiments, depending on one's other theoretical commitments.

Both mereology and Fine's theory are incomplete as accounts of parthood. Contra Lewis classical mereology is in need of considerable supplementation, and we have demonstrated a need for adding not only topological concepts but also ones pertaining to function. To the best of my knowledge this has not yet been attempted in the literature, and would be a valuable development of the theory. As Steinberg's analysis of oceanic space suggests, the same is equally true of the theory of boundaries. Fine's theory of embodiment is also incomplete insofar as he applies it only to artefacts such as cars, and it could be usefully developed by exploring whether it applies equally well to natural kinds, for example, and what characteristic patterns of embodiment they may require.

A further useful development in ontology arising from this thesis would be the development of a formal theory of representation. We have seen that representation should properly be considered to be formal in Poli's sense, yet there is no rigorous analysis of the concept available. Developing a formal theory of representation, by analogy with mereology's analysis of parthood, would shed light on the meaning of this difficult concept. In particular, such a theory would be invaluable in determining the relationship between uses of the term representation in a variety of contexts including the arts and the natural sciences. It may be, for example, that the same concept is at play in all cases, but is restricted in scope by domain-specific qualifiers. Alternatively, demonstrating that the uses of representation in these and/or other contexts are incommensurable would be a valuable result, helping to prevent further equivocation.

Finally, in chapter six we canvassed a radically externalist identity theory of mind. While the intention was not to seriously suggest the theory as a novel account of the mind, only to provide a foil to illustrate how Fodor and Millikan's semantics do not require mental representations, it may be that the rough sketch drawn here can be fleshed out to a plausible theory. There are significant hurdles to overcome: as with all reductive theories we would have to explain how talk which appears to be about one thing is really about something else; also it is unclear what processes might be able to explain inference and

patterns of reasoning. Nevertheless we have seen that the theory is parsimonious, and while it is counter to many philosophers' intuitions, that may not be such a bad thing.

Bibliography

Adams, F. & Aizawa, K. 2010. Causal Theories of Mental Content *The Stanford Encyclopedia of Philosophy* http://plato.stanford.edu/archives/spr2010/entries/content-causal/.

Albertazzi, L. 1996. Material and Formal Ontology. In *Formal Ontology* ed. P. Simons, & R. Poli, 211-47. Dordrecht: Kluwer.

Arlig, A. 2008. Medieval Mereology. *The Stanford Encyclopedia of Philosophy* http://plato.stanford.edu/archives/win2008/entries/mereology-medieval/.

Armstrong, D. 1997. *A World of States of Affairs*. Cambridge: Cambridge University Press.

Aydede, M. 2010. The Language of Thought Hypothesis. *The Stanford Encyclopedia of Philosophy* http://plato.stanford.edu/archives/fall2010/entries/language-thought.

Azzouni, J. 1998. On "On What There Is". *Pacific Philosophical Quarterly* 79 (1): 1–18.

Azzouni, J. 2007. Ontological Commitment in the Vernacular. *Noûs* 41 (2): 204–226

Baker, L. R. 1997. Why Constitution Is Not Identity. *Journal of Philosophy* 94: 599-621.

Baker, L. R. 1999. What is this thing called 'Commonsense Psychology'? *Philosophical Explorations* 2: 3-19.

Bar-Hillel, M. 1977. The Base Rate Fallacy in Probability Judgments. *Acta Psychologica* 44, 211-233.

Baron-Cohen, S. 1989. The Autistic Child's Theory of Mind: A Case of Specific Developmental Delay. *Journal of Child Psychology and Psychiatry* 30: 285 - 98.

Barth, H., Kanwisher, N., & Spelke, E. S. 2003. Construction of Large Number Representations in Adults. Cognition 86, 3: 201 – 221.

Baxter, D. L. M. 1988. Identity in the Loose and Popular Sense. *Mind* 97: 575-82.

Bermudez, J. L. 2005. *Philosophy of Psychology: A Contemporary Introduction*. London: Routledge.

Bermudez, J. L. 2010. Nonconceptual Mental Content. *The Stanford Encyclopedia of Philosophy* http://plato.stanford.edu/archives/spr2010/entries/content-nonconceptual/.

Bickle, J. 2008. Multiple Realizability. *The Stanford Encyclopedia of Philosophy* http://plato.stanford.edu/archives/fall2008/entries/multiple-realizability.

Block, N. 1998. Conceptual Role Semantics. In *The Routledge Encyclopedia of Philosophy*, ed. *E.* Craig. London: Routledge. Also available at

http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/ConceptualRoleSemantics.html.

Bogdan, R. J. (ed.) 1996. *Belief.* Oxford: Oxford University Press.

Botterill, G. & Carruthers, P. 1999. *The Philosophy of Psychology.* Cambridge: Cambridge University Press.

Brown, C. 2002. Narrow Mental Content. *The Stanford Encyclopedia of Philosophy* http://plato.stanford.edu/entries/content-narrow/.

Brentano, F. 1995. *Psychology from an Empirical Standpoint.* Originally published in 1874; English edition edited by L. McAlister, London: Routledge and Kegan Paul 1973; reprinted with an introduction by Peter Simons, London: Routledge 1995.

Burkhardt, H. & Dufour, C. A. 1991. Part-whole I: History. In *Handbook of Metaphysics and Ontology, ed.* Burkhard, H. & Smith, B., 663-73. Munich: Philosophia Verlag.

Burge, T. 1977. Individualism and the Mental. *Midwest Studies in Philosophy* 4 (1): 73-122.

Burge, T. (1982) 'Other Bodies' In *Thought and Object*, ed. A. Woodfield. Oxford: Oxford University Press.

Buttelmann, D., Carpenter, M., Call, J. & Tomasello, M. 2008. Rational Tool Use and Tool Choice in Human Infants and Great Apes. *Child Development* 79 (3): 609–26.

Byrne, R. M. & Russon, A. E. 1998. Learning by Imitation: A Hierarchical Approach. *Behavioural and Brain Sciences* 21: 667-721.

Cameron, R. 2009. Mereological Essentialism. In *Handbook of Mereology,* ed. H. Burkhardt, J. Seibt, & G. Imaguire. Munich: Philosophia Verlag. Also available at http://www.personal.leeds.ac.uk/~phlrpc/Mereological%20Essentialism.pdf.

Cantor, G. 1899. Letter to Dedekind. In *From Frege to Godel: A Sourcebook in Mathematical Logic, 1879-1931,* ed. J. Van Heijenoort. Cambridge, MA: Harvard University Press.

Carruthers, P. & Smith, P. (eds.) 1995. *Theories of Theories of Mind*. Cambridge: Cambridge University Press.

Casati, R. & Varzi, A. 1994. *Holes and Other Superficialities.* Cambridge, MA: MIT Press.

Casati, R. & Varzi, A. 1999. *Parts and Places: The Structures of Spatial Representation*. Cambridge, MA: MIT Press.

Chalmers, D. 1996. *The Conscious Mind.* Oxford: Oxford University Press

Chalmers, D. 2004. The Representational Character of Experience. In *The Future for Philosophy*, ed. B. Leiter. Oxford: Oxford University Press. Also available at http://consc.net/papers/representation.pdf.

Chamlers, D. 2006. Probability and Propositions. Available at http://consc.net/papers/probability.pdf.

Chiang, W. & Wynn, K. 2000. Infants' Tracking of Objects and Collections. *Cognition* 77 (3): 169-95.

Chisholm, R. M. 1956. *Perceiving: a Philosophical Study.* Ithaca, NY: Cornell University Press.

Chisholm, R. M. 1975. Mereological Essentialism: Further Considerations. *Review of Metaphysics* 28: 477-484.

Churchland, P. M. 1981. Eliminative Materialism and the Propositional Attitudes. *Journal of Philosophy* 78: 67-90.

Churchland, P. M. 1988. *Matter and Consciousness, Revised Edition.* Cambridge, MA: MIT Press.

Clapin, H. (ed.) 2002a. *Philosophy of Mental Representation*. Oxford: Oxford University Press.

Clapin, H. 2002b. Tacit Representation in Functional Architecture. In *Philosophy of Mental Representation,* ed. H. Clapin.

Clark, A. 1987. From Folk Psychology to Naïve Psychology. *Cognitive Science* 11: 139-54.

Clark, A., & Chalmers, D. 1998. The Extended Mind. *Analysis* 58 (1): 7-19.

Clarke, A. C. 1945. Peacetime Uses for V2: V2 for Ionosphere Research? *Wireless World* 51: 2.

Cole, J, 2005. On the Relation of the Body Image to Sensation and its Absence. In *Body Image and Body Schema: Interdisciplinary Perspectives on the Body*, ed. H. de Preester & V. Knoeckaert. Amsterdam: John Benjamins.

Correia, F. 2005. *Existential Dependence and Cognate Notions.* Munich: Philosophia Verlag.

Crane, T. 1995. *The Mechanical Mind.* London: Routledge.

Crane, T. 2006. Brentano's Concept of Intentional Inexistence. In *The Austrian Contribution to Philosophy*, ed. M. Textor, 20-35. London: Routledge.

Cresswell, M. J. 2004. Adequacy Conditions for Counterpart Theory. *Australasian Journal of Philosophy* 82 (1): 28-41.

Cresswell, M. J. 2005. Review of 'Ways a World Might Be: Metaphysical and Anti-Metaphysical Essays' by Stalnaker, R. *Australasian Journal of Philosophy* 83 (3): 434-7.

Cruse, D. A. 1979. On the Transitivity of the Part-Whole Relations. *Journal of Linguistics* 15: 29-38.

Cummins, R. 1975. Functional Analysis. *The Journal of Philosophy* 72 (20): 741-65.

Cummins, R. 1981. *Meaning and Mental Representation.* Cambridge, MA: MIT Press.

Cummins, R. 1996. *Representations, Targets and Attitudes*. Cambridge, MA: MIT Press.

Cummins, R. 2002. *Neo*-Teleology. In *Functions: New Essays in the Philosophy of Psychology and Biology,* ed. R. Cummins, A. Ariew & M. Perlman, 157-72. Oxford: Oxford University Press.

Cummins, R., Ariew, A., & Perlman, M. (eds.) 2002. Functions: New Essays in the Philosophy of Psychology and Biology. Oxford: Oxford University Press.

Davidson, D. 1987. Knowing One's Own Mind. *Proceedings of the American Philosophical Association* 61: 441-58.

Davies, M. & Stone, T. (eds.) 1995a. *Folk Psychology and the Theory of Mind Debate*. Oxford: Blackwell.

Davies, M. & Stone, T. (eds.) 1995b. *Mental Simulation: Evaluations and Applications*. Oxford: Blackwell.

de Preester, H. & Knockaert, V. (eds.) 2005. *Body Image and Body Schema: Interdisciplinary Perspectives on the Body.* Amsterdam: John Benjamins.

de Vignemont, F. 2006. A Review of Shaun Gallagher, *How the Body Shapes the Mind. Psyche* 12 (1) *http://journalpsyche.org/ojs-2.2/index.php/psyche/article/viewFile/2759/2625*.

Dennett, D. 1982. Beyond Belief. In *Thought and Object: Essays in Intentionality*, ed. A. Woodfield. Oxford: Oxford University Press.

Dennett, D. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.

Dennett, D. 1991. *Consciousness Explained.* Boston: Little, Brown & Co.

Dennett, D. 1995. Darwin's Dangerous Idea: Evolution and the Meanings of Life*. London: Penguin Books.*

Dretske, F. 1981. *Knowledge and the Flow of Information.* Cambridge, MA: MIT Press.

Dretske, F. 1986. Misrepresentation. In *Belief*, ed. R. J. Bogdan. Oxford: Oxford University Press.

Dretske, F. 1988. *Explaining Behaviour.* Cambridge, MA: MIT Press.

Dretske, F. 1990. Reply to Reviewers. *Philosophy and Phenomenological Research* 1:4, 819-39.

Dretske, F. 1995. *Naturalizing the Mind.* Cambridge, MA: MIT Press.

Dretske, F. 1998. An Interview with Fred Dretske. *The Dualist* 5 http://www.stanford.edu/group/dualist/vol5/pdfs/dretske.pdf.

Dretske, F. 2006. Representation and Self-Knowledge. In *Teleosemantics: New Philosophical Essays,* ed. G. Macdonald & D. Papineau. Oxford: Oxford University Press.

Evans, G. 1982. *The Varieties of Reference*. Oxford: Oxford University Press.

Fara, D. G. 2007. Counterparts Within Actuality. From the *2nd Online Philosophy Conference* http://experimentalphilosophy.typepad.com/2nd_annual_online_philoso/files/delia_graff_fara.pdf .

Fine, K. 1995. Part-Whole. In *The Cambridge Companion to Husserl*, ed. B. Smith & D. W. Smith, 463-85. Cambridge: Cambridge University Press.

Fine, K. 1999. Things and Their Parts. *Midwest Studies in Philosophy* 23 (1): 61-74.

Fodor, J. 1975. *The Language of Thought.* Cambridge, MA: Harvard University Press.

Fodor, J. 1983. *The Modularity of Mind.* Cambridge, MA: MIT Press.

Fodor, J. 1985. Fodor's Guide to Mental Representation: The Intelligent Auntie's Vade-Mecum. *Mind* 373: 76-100.

Fodor, J. 1987. *Psychosemantics.* Cambridge, MA: MIT Press.

Fodor, J. 1990a. *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Fodor, J. 1990b. Information and Representation. In *Information, Language, and Cognition,* ed. P. Hanson, 175-90. Vancouver: University of British Columbia.

Fodor, J. 1994. *The Elm and the Expert.* Cambridge, MA: MIT Press.

Fodor, J. 1998a. *Concepts: Where Cognitive Science Went Wrong*. Oxford: Oxford University Press.

Fodor, J. 1998b. *In Critical Condition: Polemical Essays on Cognitive Science and the Philosophy of Mind.* Cambridge, MA: MIT Press.

Fodor, J. 2003. *Hume Variations*. Oxford: Oxford University Press.

Fodor, J. 2008. *LOT 2: The Language of Thought Revisited.* Oxford: Oxford University Press.

Frackowiak, R. S. J. 2004. *Human Brain Function*. California: Elsevier.

Frege, G. 1956. The Thought: A Logical Inquiry. *Mind* 65 (259): 289-311.

Frith, U. & Happé, F. 1999. Theory of Mind and Self-consciousness: What is it Like to be Autistic? *Mind and Language* 14: 1-22.

Gallagher, S. 2001. The Practice of Mind: Theory, Simulation or Interaction? *Journal of Consciousness Studies* 5 (7): 83-108.

Gallagher, S. 2005a. *How the Body Shapes the Mind.* Oxford: Oxford University Press.

Gallagher, S. 2005b. Dynamic Models of Body Schematic Processes. In *Body Image and Body Schema: Interdisciplinary Perspectives on the Body*, ed. H. de Preester & V. Knoeckaert. Amsterdam: John Benjamins.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. 1996. Action Recognition in the Premotor Cortex. *Brain* 119: 593 – 609.

Gallese, V. & Goldman, A. 1998. Mirror Neurons and the Simulation Theory of Mind-Reading. *Trends in Cognitive Sciences* 2: 493 – 501.

Gallese, V., Keysers, C., & Rizzolatti, G. 2004. A Unifying View of the Basis of Social Cognition. *Trends in Cognitive Sciences* 8: 396 – 403.

Garbacz, G. 2007. A First Order Theory of Functional Parthood. *Journal of Philosophical Logic* 36: 309-37.

Gergely, G., Bekkering, H. & Kiraly, I. 2002. Rational Imitation in Preverbal Infants. *Nature* 14: 415.

Gibson, J. J. 1966. *The Senses Considered as Perceptual Systems.* London: George Allen and Unwin.

Goldman, A. 1995. Interpretation Psychologized. In *Folk Psychology and the Theory of Mind Debate*, ed. M. Davies & T. Stone. Oxford: Blackwell.

Gopnik, A. & Wellman, H. 1995. Why the Child's Theory of Mind Really *Is* a Theory. In *Folk Psychology and the Theory of Mind Debate*, ed. M. Davies & T. Stone. Oxford: Blackwell.

Gordon, R. 1995a. Folk Psychology as Simulation In *Folk Psychology and the Theory of Mind Debate*, ed. M. Davies & T. Stone. Oxford: Blackwell.

Gordon, R. 1995b. The Simulation Theory: Objections and Misconceptions. In *Folk Psychology and the Theory of Mind Debate*, ed. M. Davies & T. Stone. Oxford: Blackwell.

Gordon, R. 2000. Sellars's Ryleans Revisited. *Protosociology* 14.

Gouteux, S. & Spelke, E. S. 2001. Children's Use of Geometry and Landmarks to Reorient in an Open Space. *Cognition* 81 (2): 119-48.

Grice, H. P. 1957. Meaning. *The Philosophical Review* 66: 377-88.

Grimmitt, M. (ed.) 2000. *Pedagogies of Religious Education: Case Studies in the Research and Development of Good Pedagogic Practice in RE.* Great Wakering: McCrimmons.

Guarino, N. 1995. Formal Ontology, Conceptual Analysis and Knowledge Representation. *International Journal of Human and Computer Studies* 43 (5-6): 625-40. Also available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.29.3373.

Harman, G. 1990. The Intrinsic Quality of Experience. In *Philosophical Perspectives 4*, ed. J. Tomberlin. Atascadero: Ridgeview.

Harris, P. 1995. From Simulation to Folk Psychology: The Case for Development. In *Folk Psychology and the Theory of Mind Debate*, ed. M. Davies & T. Stone. Oxford: Blackwell.

Harte, V. 2002. *Plato on Parts and Wholes: The Metaphysics of Structure.* Oxford: Clarendon.

Hazen, A. 1979. Counterpart-theoretic Semantics for Modal Logic. *Journal of Philosophy* 76: 319-38.

Heal, J. 1995. Replication and Functionalism. In *Folk Psychology and the Theory of Mind Debate*, ed. M. Davies & T. Stone. Oxford: Blackwell.

Hendry, R. F. 2008. Two Conceptions of the Chemical Bond. *Philosophy of Science* 75: 909-20.

Hobbs, J. R. and Moore, R. C. (eds.) 1985. *Formal Theories of the Common-Sense World.* Norwood: Ablex.

Holden, T. 2004. *The Architecture of Matter: Galileo to Kant*. Oxford: Clarendon.

Horgan, T. 1993. On What There Isn't. *Philosophy and Phenomenological Research* 53: 693-700.

Hornby, M. & Peach, J. M. 1993. *Foundations of Organic Chemistry*. Oxford: Oxford University Press.

Hunter, G. & Seager, W. 1981. The Discreet Charm of Counterpart Theory. *Analysis* 41: 73-76.

Huntley-Fenner, G., Carey, S. & Solimando, A. 2002. Objects are Individuals but Stuff Doesn't Count: Perceived Rigidity and Cohesiveness Influence Infants' Representations of Small Groups of Discrete Entities. *Cognition* 85: 203-21.

Hurley, S. L. 1998. *Consciousness in Action.* Cambridge, MA: Harvard University Press.

Hurley, S. L. 2010. Varieties of Externalism. In *The Extended Mind*, ed. R. Menary. London: Ashgate.

Husserl, E. 2001. *Logical Investigations.* Trans. J. N. Findlay. London and New York: Routledge.

Jackson, F. 1982. Epiphenomenal Qualia. *Philosophical Quarterly* 32: 127-36. Also available at http://instruct.westvalley.edu/lafave/epiphenomenal_qualia.html.

Jackson, F. 2003. Mind and Illusion. *Minds and Persons: Royal Institute of Philosophy Supplement 53*: 251-71. Also available at http://consciousness.anu.edu.au/jackson/mindillusion.pdf.

Johansson, I. 2004. On the Transitivity of the Parthood Relations. In *Relations and Predicates*, ed. H. Hochberg & K. Mulligan. Frankfurt: Ontos Verlag. Also available at http://hem.passagen.se/ijohansson/function2.PDF

Johansson, I. 2006. Formal Mereology and Ordinary Language – A Reply to Varzi. *Applied Ontology* 1 (2): 157-61. Also available at http://hem.passagen.se/ijohansson/information5.pdf

Johnson, S. C. 2000. The Recognition of Mentalistic Agents in Infancy. *Trends in Cognitive Sciences* 4*:* 22 – 28.

Johnston, M. 2006. Hylomorphism. *The Journal of Philosophy* 103 (12): 652-98.

Koslicki, K. 2001. The Crooked Path from Vagueness to Four-Dimensionalism. *Philosophical Studies* 114: 107-34.

Koslicki, K. 2007. Towards a Neo-Aristotelian Mereology. *Dialectica* 61 (1): 127–59.

Kripke, S. A. 1972. *Naming and Necessity.* In *Semantics of Natural Languages*, ed. D. Davidson & G. Harman, 253-355. Dordrecht: Reidel.

Kuhn, S. 2009. Prisoner's Dilemma. *The Stanford Encyclopedia of Philosophy* http://plato.stanford.edu/archives/sum2009/entries/prisoner-dilemma/.

Kuratowski, K. 1930. Sur le Probleme des Courbes Gauches en Topologie. *Fundamenta Mathematicae* 15: 271-83.

Lau, J. & Deutsch, M. 2009. Externalism about Mental Content. *The Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/archives/sum2009/entries/content-externalism/.

Leonard, H. S. & Goodman, N. 1940. The Calculus of Individuals and Its Uses. *Journal of Symbolic Logic* 5: 45-55.

Leśniewski, S. 1916. Foundations of the General Theory of Sets I. Trans. D. I. Barnett. In his *Collected Works Vol. 1*, ed. S. J. Surma, J. Srzednicki, D. I. Barnett, and F. V. Rickey, 129-73. Dordrecht: Kluwer.

Lewis, D. K. 1968. Counterpart Theory and Quantified Modal Logic. *Journal of Philosophy* 65: 113-26. Reprinted in his *Philosophical Papers, Vol. 1.* Oxford: Oxford University Press.

Lewis, D. K. 1972. Psychophysical and Theoretical Identifications. *Australasian Journal of Philosophy* 50: 249-58.

Lewis, D. K. 1986. *On the Plurality of Worlds.* Oxford: Blackwell.

Lewis, D. K. 1988. Rearrangement of Particles: Reply to Lowe. *Analysis* 48: 65-72.

Lewis, D. K. 1991. *Parts of Classes.* Oxford: Blackwell.

Lewis, D. 1994. Reduction of Mind. In *A Companion to Philosophy of Mind*, ed. S. Guttenplan. Oxford: Blackwell.

Lipton, J., & Spelke, E. S. 2000. Infants' Discrimination of Large Numbers of Sounds. Unpublished manuscript referenced in Spelke 2000.

Loar, B. 1988. *Social Content and Psychological Content.* In *Contents of Thought*, ed. R. H. Grimm & D. D. Merrill. Tucson, AZ: University of Arizona.

Lowe, E. J. 1986. On a Supposed Temporal/Modal Parallel. *Analysis* 46: 195-7.

Lowe, E. J. 1989. *Kinds of Being: A Study of Individuation, Identity and the Logic of Sortal Terms*. Oxford: Blackwell.

Lowe, E. J. 1998. Entity, Identity and Unity. *Erkenntnis* 48: 191-208.

Lowe, E. J. 2006. *The Four-Category Ontology.* Oxford: Oxford University Press.

Lowe, E. J. 2010. Ontological Dependence. *The Stanford Encyclopedia of Philosophy* http://plato.stanford.edu/archives/spr2010/entries/dependence-ontological/.

Mann, W. & Varzi, A. C. (eds.) 2006. *Parts and Wholes.* Special issue of *The Journal of Philosophy*, 103.

Markosian, N. 1998. Brutal Composition. *Philosophical Studies* 92: 211-29.

Markosian, N. 2007. Restricted Composition. In *Contemporary Debates in Metaphysics,* ed. J. Hawthorne, T. Sider & D. Zimmerman. Oxford: Blackwell.

Marr, D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* San Francisco: W. H. Freeman.

Matthews, R. J. 2007. *The Measure of Mind.* Oxford: Oxford University Press.

McCall, S., & Lowe, E. J. 2003. 3D/4D Equivalence, the Twins Paradox and Absolute Time. *Analysis* 63: 114–23.

McDowell, J. 1984. *De Re* Senses *Philosophical Quarterly* 34: 283-94.

McGinn, C. 1989. *Mental Content.* Oxford: Blackwell.

McLaughlin, B. (ed.) 1991. *Dretske and His Critics.* Boston: Blackwell.

Melia, J. 2007. Response to Counterparts Within Actuality. From the *2nd Online Philosophy Conference* http://experimentalphilosophy.typepad.com/2nd_annual_online_philoso/files/melia_on_delia.doc.

Menary, R. (ed.) 2010. *The Extended Mind*. Cambridge, MA: MIT Press.

Menger, K. 1940. Topology Without Points. *Rice Institute Pamphlets* 27: 80-107.

Michotte, A. 1963. *The Perception of Causality.* Trans. T. R. Miles & E. Miles. New York: Basic Books.

Millikan, R. G. 1984. *Language, Thought, and Other Biological Categories.* Cambridge, MA: MIT Press.

Millikan, R. G. 1989. In Defense of Proper Functions. *Philosophy of Science* 56: 288-302.

Millikan, R. G. 1995. Biosemantics. In her *White Queen Psychology and Other Essays for Alice*. Cambridge, MA: MIT Press.

Millikan, R. G. 1996. Pushmi-pullyu Representations. In *Philosophical Perspectives vol. IX*, ed. J. Tomberlin, 185-200. Atascadero CA: Ridgeview Publishing.

Millikan, R. G. 2000a. *On Clear and Confused Ideas.* Cambridge: Cambridge University Press. Also available at http://www.philosophy.uconn.edu/department/millikan/clearct.htm.

Millikan, R. G. 2000b. Representations, Targets and Attitudes. *Philosophy and Phenomenological Research* 60: 103-11.

Millikan, R. G. 2000c. Reading Mother Nature's Mind. In *Dennett's Philosophy,* ed. D. Ross, A. Brook & D. Thompson. Cambridge, MA: MIT Press.

Millikan, R. G. 2002a. Biofunctions: Two Paradigms. In *Functions: New Readings in the Philosophy of Psychology and Biology,* ed. R. Cummins, A. Ariew and M. Perlman, 113-43. Oxford: Oxford University Press.

Millikan, R. G. 2002b. *Mental Content, Teleological Theories of.* In *Encyclopedia of Cognitive Science*, ed. L. Nagel. London: Macmillan. Also available at http://www.philosophy.uconn.edu/department/millikan/teleocnt.pdf

Millikan, R. G. 2004. *Varieties of Meaning: The Jean Nicod Lectures 2002.* Also available at http://www.philosophy.uconn.edu/department/millikan/contents.htm.

Millikan, R. G. 2005. *Language: A Biological Model.* Oxford: Clarendon. Also available at http://www.philosophy.uconn.edu/department/millikan/langct.htm.

Millikan, R. G. 2008. Biosemantics. In *The Oxford Handbook of Philosophy of Mind*, ed. B. McLaughlin, A. Beckermann, & S. Walter. Oxford: Oxford University Press.

Mishara, A. L. 2005. Body Self and its Narrative Representation in Schizophrenia: Does the Body Schema Concept Help Establish a Core Deficit? In *Body Image and Body Schema: Interdisciplinary Perspectives on the Body*, ed. H. de Preester & V. Knoeckaert. Amsterdam: John Benjamins.

Molina, M., Van de Walle, G. A., Condry, K. & Spelke, E. S. 2004. The Animate-Inanimate Distinction in Infancy: Developing Sensitivity to Constraints on Human Actions. *Journal of Cognition and Development* 5 (4): 399–426.

Moore, G. E. 1959. A Defence of Commonsense. In his *Philosophical Papers, 32-59.* London: George Allen and Unwin.

Moran, D. 1996. Brentano's Thesis. *Proceedings of the Aristotelian Society* 70: 1-27

Morton, A. 1996. Folk Psychology is Not a Predictive Device. *Mind* 105 (417): 119-37.

Needham, P. 2003. Chemical Substances and Intensive Properties. *Annals of the New York Academy of Sciences* 988: 99-113.

Nida-Rümelin, M. 2009. Qualia: The Knowledge Argument. *The Stanford Encyclopedia of Philosophy* http://plato.stanford.edu/archives/win2009/entries/qualia-knowledge.

Nuccetelli, S. (ed.) 2003. *New Essays on Semantic Externalism and Self-Knowledge.* Cambridge, MA: MIT Press.

Olsen, E. T. 1995. Why I Have no Hands. *Theoria* 61: 182-97.

Papineau, D. 2009. Ontological Naturalism. *The Stanford Encyclopedia of Philosophy* http://plato.stanford.edu/archives/spr2009/entries/naturalism/.

Parsons, J. 2004. Dion, Theon and DAUP. *Pacific Philosophical Quarterly* 85 (1): 85-91.

Pessin, A. & Goldberg, S. (eds.) 1996. *The Twin Earth Chronicles.* London: M.E. Sharpe.

Pinker, S. 2002. *The Blank Slate: The Modern Denial of Human Nature.* New York: Viking.

Place, U. T. 1956. Is Consciousness a Brain Process? *British Journal of Psychology* 47: 44-50

Plantinga, A. 1974. *The Nature of Necessity.* New York: Oxford University Press.

Poli, R. 1993. Husserl's Conception of Formal Ontology. *History and Philosophy of Logic* 14: 1-14.

Poli, R. 2003. Descriptive, Formal and Formalized Ontologies. In *Husserl's Logical Investigations Reconsidered,* ed. D. Fisette, 183-210. Dordrecht: Kluwer. Also available at http://www.formalontology.com/essays/descriptive-ontologies.pdf.

Putnam, H. 1967. The Nature of Mental States. In *Art, Mind, and Religion*, ed. W. H. Capitan, & D. D. Merrill. Pittsburgh: Pittsburgh University Press.

Putnam, H. 1975. The Meaning of 'Meaning'. In *Language, Mind and Knowledge, Minnesota Studies in the Philosophy of Science, VII*, ed. K. Gunderson. Minneapolis: University of Minnesota Press. Reprinted in Putnam, H. 1975. *Mind, Language and Reality: Philosophical Papers, Volume 2,* 215-271. Cambridge: Cambridge University Press.

Quine, W. V. O. 1951. On What There Is. In his *From A Logical Point of View, 1-19.* Cambridge, MA: Harvard University Press.

Quine, W.V.O. 1960. *Word and Object.* Cambridge, MA: MIT Press.

Quine, W. V. O. 1981. Worlds Away. In his *Theories and Things.* Cambridge, MA: Harvard University Press.

Quine, W. V. O. 1974. *The Roots of Reference*. La Salle, IL: Open Court.

Raley, Y. 2005. Ontological Naturalism. *Pacific Philosophical Quarterly* 86 (2): 284-94.

Ramachandran, M. 1989. An Alternative Translation Scheme for Counterpart Theory. *Analysis* 49: 131-41.

Ramachandran, M. 1990a. Contingent Identity in Counterpart Theory. *Analysis* 50: 163-6.

Ramachandran, M. 1990b. Unsuccessful Revisions of CCT. *Analysis* 50: 173-7.

Ramachandran, V. S. & Hubbard, E. M. 2003. More Common Questions about Synesthesia. *Scientific American* http://www.scientificamerican.com/article.cfm?id=more-common-questions-abo-2003-04-14.

Ramsey, T. Z. 2006. How the Body Shapes the Mind: An Interview with Shaun Gallagher. *Science & Consciousness Review* http://pegasus.cc.ucf.edu/~gallaghr/scrInterview06.html.

Ratcliffe, M. 2006. Folk Psychology is Not Folk Psychology. *Phenomenology and the Cognitive Sciences* 5 (1): 31-52.

Ratcliffe, M. 2007. *Rethinking Commonsense Psychology: A Critique of Folk Psychology, Theory of Mind and Simulation*. Basingstoke: Palgrave Macmillan.

Ravenscroft, I. 2004. Folk Psychology as a Theory. *The Stanford Encyclopedia of Philosophy* http://plato.stanford.edu/archives/win2004/entries/folkpsych-theory/.

Rescher, N. 1955. Axioms for the Part Relation. *Philosophical Studies* 6: 8-11.

Rizzolatti, G.& Craighero, L. 2001. The Mirror Neuron System. *Annual Review of Neuroscience* 27: 169-92.

Rosen, G. & Dorr, C. 2003. Composition as a Fiction. In *The Blackwell Guide to Metaphysics*, ed. R. M. Gale. Oxford: Blackwell.

Rosenberg, R. D. & Carey, S. 2006. Infants' Indexing of Objects vs. Non-Cohesive Entities. Paper presented at the *Annual Meeting of the XVth Biennial International Conference on Infant Studies,* Westin Miyako, Kyoto, Japan, Jun 19 2006.

Rowlands, M. 2003. *Externalism: Putting Mind and World Back Together Again*. Chesham: Acumen.

Rowlands, M. 2006. *Body Language: Representing in Action*. Cambridge, MA: MIT Press.

Sanford, D. 1993. The Problem of the Many, Many Composition Questions, and Naive Mereology. *Noûs* 27: 219-28.

Sanford, D. H. 2003. Fusion Confusion. *Analysis* 63: 1-4.

Sauvagnat, F. 2005. Body Structure in Psychotic and Autistic Children. In *Body Image and Body Schema: Interdisciplinary Perspectives on the Body*, ed. H. de Preester & V. Knoeckaert. Amsterdam: John Benjamins.

Schiffer, S. 1981. Truth and the Theory of Content. In *Meaning and Understanding*, ed. H. Parret & J. Bouvaresse. Berlin: Walter de Gruyter.

Schmuckler, M. A. & Li, N. S. 1998. Looming Responses to Obstacles and Apertures: The Role of Accretion and Deletion of Background Texture. *Psychological Science* 9 (1): 49-52.

Schmuckler, M. A., Collimore, L. M. & Dannelimmer, J. L. 2007. Infants' Reaction to Object Collision on Hit and Miss Trajectories. *Infancy* 12 (1): 105-18.

Sellars, W. 1997. *Empiricism and the Philosophy of Mind.* Cambridge, MA: Harvard University Press.

Segal, G. 2000. *A Slim Book About Narrow Content.* Cambridge, MA: MIT Press.

Segal, G. 2005. Intentionality. In *The Oxford Handbook of Contemporary Philosophy*, ed. F. Jackson & M. A. Smith. Oxford: Oxford University Press.

Shaffer, D. R. & Kipp, K. 2009. *Developmental Psychology: Childhood and Adolescence.* Belmont, CA: Wadsworth.

Shanahan, M. P. 1996. Folk Psychology and Naive Physics. In *Connectionism, Concepts and Folk Psychology: The Legacy of Alan Turing, vol. 2,* ed. A. Clark & P. J .R. Millican. Oxford: Oxford University Press.

Shapiro, L. 2000. Multiple Realizations. *The Journal of Philosophy* 97 (12): 635-54.

Sider, T. 2001. *Four-Dimensionalism: An Ontology of Persistence and Time.* Oxford: Oxford University Press.

Sider, T. 2007a. Parthood. *Philosophical Review* 116: 51-91.

Sider, T. 2007b. Temporal Parts. In *Contemporary Debates in Metaphysics*, ed. J. Hawthorne, T. Sider & D. Zimmerman. Oxford: Blackwell.

Simons, P. 1987. *Parts: A Study in Ontology.* Oxford: Clarendon Press.

Simons, P. 1995. Introduction. In Brentano, F. *Psychology from an Empirical Standpoint*.

Simons, P. 2006. Real Wholes, Real Parts: Mereology without Algebra *The Journal of Philosophy* 103 (12): 597-613.

Smart, J. J. C. 1959. Sensations and Brain Processes. *Philosophical Review* 68: 141-56.

Smart, J. J. C. 2008. The Identity Theory of Mind. *The Stanford Encyclopedia of Philosophy* http://plato.stanford.edu/archives/fall2008/entries/mind-identity.

Smith, B. 1989. Husserl: Logic and Formal Ontology. In *Husserl's Phenomenology: A Textbook*, ed. J. N. Mohanty, & W. McKenna, 29-67. Lanham: University Press of America. Also available at http://ontology.buffalo.edu/smith/articles/lfo.html.

Smith, B. 1994. *Austrian Philosophy.* LaSalle, IL: Open Court.

Smith, B. 1996a. Boundaries: An Essay in Mereotopology. In *The Philosophy of Roderick Chisholm*, ed. L. Hahn, 534-61. LaSalle, IL: Open Court.

Smith, B. 1996b. Mereotopology: A Theory of Parts and Boundaries. *Data and Knowledge Engineering* 20: 287-303.

Smith, B. & Casati, R. 1994. Naïve Physics: An Essay in Ontology. *Philosophical Psychology* 7 (2): 225-44.

Smith, B. & Grenon, P. 2004. The Cornucopia of Formal-Ontological Relations. *Dialectica* 58 (3): 279-96.

Smith, B. & Varzi, A. 2000. Fiat and Bona Fide Boundaries. *Philosophy and Phenomenological Research* 60 (2): 401–20.

Smith, B. & Varzi, A. 2001. Environmental Metaphysics. In *Metaphysics in the Post-Metaphysical Age*, ed. P. Simons & U. Meixner. Vienna: obv&hpt.

Sober, E. 1984. *The Nature of Selection.* Cambridge, MA: MIT Press.

Sober, E. 2008. Fodor's *Bubbe Meisse* Against Darwinism. *Mind and Language* 23 (1): 42-9.

Soteriou, M. 2009. The Disjunctive Theory of Perception. *The Stanford Encyclopedia of Philosophy* http://plato.stanford.edu/archives/fall2009/entries/perception-disjunctive.

Spelke, E. S. 1990. Principles of Object Perception. *Cognitive Science* 14: 29-56.

Spelke, E. 2000. Core Knowledge. *American Psychologist* 55: 1233-43.

Spelke, E.S. 2004. Core Knowledge. In *Attention and Performance, vol. 20: Functional Neuroimaging of Visual Cognition*, ed. N. Kanwisher & J. Duncan. Oxford: Oxford University Press.

Spelke, E. S. & Kinzler, K. D. 2007. Core Knowledge *Developmental Science* 10 (1): 89-96.

Stalnaker, R. 1990. Narrow Content. In *Propositional Attitudes: The Role of Content in Logic, Language and Mind*, ed. C. A. Anderson & J. Owens. Stanford, CA: CSLI Lecture Notes.

Stalnaker, R. 1998. What Might Nonconceptual Content Be? In *Philosophical Issues*, vol. *9*, ed. E. Villanueva, 339-52. Atascadero, CA: Ridgeview.

Stamenov, M. 2005. Body Schema, Body Image, and Mirror Neurons. In *Body Image and Body Schema: Interdisciplinary Perspectives on the Body*, ed. H. de Preester & V. Knoeckaert. Amsterdam: John Benjamins.

Steinberg, P. E. 2001. *The Social Construction of the Ocean*. Cambridge: Cambridge University Press.

Stich, S. 1983. *From Folk Psychology to Cognitive Science.* Cambridge, MA: MIT Press.

Stich, S. 1994. What is a Theory of Mental Representation?' In *Mental Representation: A Reader*, ed. S. Stich & T. Warfield, 347–64. Oxford: Blackwell.

Stich, S. & Ravenscroft, I. 1994. What is Folk Psychology? *Cognition* 50: 447-68.

Stich, S. & Warfield, T. (eds.) 1994. *Mental Representation: A Reader.* Oxford: Blackwell.

Streri, A., Gentaz, E., Spelke, E. S., & Van de Walle, G. 2004. Infants' Haptic Perception of Object Unity in Rotating Displays. *The Quarterly Journal of Experimental Psychology 57A (*3): 523-38.

Tarski, A., 1929. Les Fondements de la Géométrie des Corps. *Ksiega Pamiatkowa Pierwszkego Polskiego Zjazdu Matematycznego*, supplement to *Annales de la Société Polonaise de Mathématique* 7: 29-33. Reprinted as 'Foundations of the Geometry of Solids' in Tarski, A. (1956) *Logic, Semantics, Metamathematics.* Trans. J. H. Woodger. Oxford: Clarendon Press.

Tarski, A. 1956. On The Foundations of Boolean Algebra. In his *Logic, Semantics, Metamathematics.* Trans. J. H. Woodger. Oxford: Clarendon Press.

Tuomela, R. 2002. *The Philosophy of Social Practices: A Collective Acceptance View.* Cambridge: Cambridge University Press.

Unger, P. 1979. There Are No Ordinary Objects. *Synthese* 41: 117-54.

Unger, P. 1980. Skepticism and Nihilism. *Nous* 14: 517-45.

van Inwagen, P. 1990. *Material Beings.* Ithaca, NY: Cornell University Press.

van Inwagen, P. 1994. Composition as Identity. In *Philosophical Perspectives, 8: Logic and Language*, ed. J. E. Tomberlin, 207-20. Ridgeview: Atascadero.

van Inwagen, P. 2001. The Doctrine of Arbitrary Undetached Parts. In his *Ontology, Identity and Modality: Essays in Metaphysics*. Cambridge: Cambridge University Press.

van Inwagen, P. 2006. Can Mereological Sums Change Their Parts? Journal of Philosophy 103 (12): 614-30.

Varzi, A. 2000. Mereological Commitments. *Dialectica* 54: 283-305. Also available at http://www.columbia.edu/~av72/papers/Dialectica_2000.pdf.

Varzi, A. 2002. Words and Objects. In *Individuals, Essence, and Identity, Themes of Analytic Metaphysics*, ed. A. Bottani, M. Carrara & D. Giaretta, 49-75. Dordrecht: Kluwer.

Varzi, A. 2006. A Note on the Transitivity of Parthood. *Applied Ontology* 1 (2): 141-6. Also available at http://www.columbia.edu/~av72/papers/AO_2006.pdf.

Varzi, A. 2010. Mereology. *The Stanford Encyclopedia of Philosophy.* http://plato.stanford.edu/archives/spr2010/entries/mereology/

Varzi, A. Unpublished 2006. Lowe on Vagueness and Endurance.

Vieu, L. & Aurnague, M. 2007. Part-of Relations, Functionality and Dependence. In *The Categorization of Spatial Entities in Language and Cognition*, ed. L. Vieu & M. Aurnague, 307-37. Amsterdam: John Benjamins.

Von Eckhardt, B. 1995. Folk Psychology. In *A Companion to the Philosophy of Mind*, ed. S. Guttenplen. Oxford: Blackwell.

Weisstein, E.W. Topology. *MathWorld - A Wolfram Web Resource* http://mathworld.wolfram.com/Topology.html.

Wellman, H., Cross, D., & Watson, J. 2001. Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief. *Child Development* 72: 655-84.

Wiggins, D. 1997. Sortal Concepts: A Reply to Xu. *Mind and Language* 12: 413-21.

Wiggins, D. 2001. *Sameness and Substance Renewed*. Oxford: Blackwell.

Williams, J. R. G. 2007. Eligibility and Inscrutability. *Philosophical Review* 116 (3):361-99. Also available at http://www.personal.leeds.ac.uk/~phljrgw/JRGWilliamsPhilRevEligibilityInscrutability.pdf.

Wilson, R. A. 2004. *Boundaries of the Mind: The Individual in the Fragile Sciences.* Cambridge: Cambridge University Press.

Wilson, R. A. 2005. Collective Memory, Group Minds and the Extended Mind Thesis. *Cognitive Processing* 6 (4): 227-36. Also available at http://www.arts.ualberta.ca/~raw/collectmem.pdf

Wimmer, H. & Perner, J. 1983. Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception. *Cognition* 13: 103-28.

Winston, M., Chaffin, R., & Herrmann, D. 1987. A Taxonomy of Part-Whole Relations. *Cognitive Science* 11: 417-44.

Wynn, K. 1992. Addition and Subtraction by Human Infants. *Nature* 358: 749-50.

Xu, F. 1997. From Lot's Wife to a Pillar of Salt: Evidence that Physical Object is a Sortal Concept. *Mind and Language* 12: 365-92.

Xu, F. & Spelke, E. S. 2000a. Large Number Discrimination in Infants: Evidence for Analog Magnitude Representations. Paper presented at the *International Conference on Infant Studies*, Brighton, England. Referenced in Spelke, E. S. 2000.

Xu, F. & Spelke, E. S. 2000b. Large Number Discrimination in 6-month-old Infants. *Cognition* 74: 1-11.

Yi, B-U. 1999. Is Mereology Ontologically Innocent? Philosophical Studies 93 (2): 141-60.

Zangwill, N. 1992. Variable Reduction Not Proven. *Philosophical Quarterly* 42: 214-218.