NATURAL MINDS

Maps, mental causation and virtual machines

Simon Christopher Bowes

FEBRUARY 1, 2017

THE UNIVERSITY OF SUSSEX

Thesis submitted for the degree of PhD Cognitive Science

Contents

Introduction

My project is an empirically informed investigation of the philosophical problem of mental causation, and simultaneously a philosophical investigation of the status of cognitive scientific generalisations. If there is such a thing as mental causation, that is, mental states having effects *qua* mental states, and if we can classify the mental states involved in these causes in a way useful for making predictions and giving scientific explanations, then these states will be natural kinds. The first task, then, is to show that there is an account of natural kindhood that can accommodate cognitive kinds. The second task is to say how the scientific statements made using these mental kinds are not susceptible to being reduced to statements about physical kinds, and in fact require taking into account facts at many levels of explanation, including the biological and social levels. Lastly, the case will be made for Virtual Machine Functionalism being the correct account of the relationship between cognitive states and the broader physical world.

I will claim that although there may be problems with traditional accounts of natural kinds and mental representations when it comes to contemporary cognitive science, this is no reason for thinking that those terms are not useful; we should refine rather than eliminate them. Something would be lost in our understanding if we rejected these terms and the theoretical understanding they contain, something that was present before some mistaken theoretical details came to be associated with the terms. Perhaps some terms that have been coined in the development of our understanding of the mind, such as 'qualia,' should be dispensed with, but others just need to be cleaned up.

Broadly speaking, my argument will be that squaring the widely held but somewhat contradictory intuitions of physicalism and anti-reductionism regarding mental states will require modification of two other commonly held intuitions, namely physical causal closure and supervenience.

Another way of stating my aim is in terms of defending the intuitive distinction between metaphorical and literal uses of intentional vocabulary, such as 'wanting' and 'trying,' against those who question the meaningfulness of this distinction because they take a physicalist stance on questions of consciousness. They may say the distinction is merely verbal, that there is no real difference between saying of a raindrop that it is 'trying' to get to the bottom of the window pane, and saying of a person that he or she is trying to get to the top of the mountain. Much of what follows is an attempt to describe a metaphysics that is materialist and scientific, but in which the difference between the two cases has a natural place.

The difference lies in the idea of intentionality: in the case of my desiring something, there is a mental state 'in' me that is there because it has the function of directing my actions towards bringing about the desired state of affairs. Such states are things that can be scientifically studied, and a scientific account of human action would be incomplete if it did not refer to such states. In the case of the water drop, there are no similar states without which the scientific understanding of water droplet action would be incomplete. The temptation to elide the distinction between intentional and non-intentional descriptions is based on a belief that since all causation is physical, there is no meaningful distinction between the kinds of causes that makes water drops drop, and those that make climbers climb. This results from the fact that it is sometimes felt that reference to such things as personal agency in intentional explanations of action is to allow in a disagreeable form of dualism. I will argue that a complete, physicalistic, scientific account of human behaviour must include reference to irreducible, mental kinds, such as beliefs and desires.

The form of the argument follows the content, with natural kinds at the centre of the web of concepts that form our understanding of mind and its place in the natural world. The starting point is simple folk explanations of human actions, like, 'He ate the apple because…' followed by a set of conditions including combinations of beliefs and desires that together constitute sufficient reasons for eating the apple. Many would say such purported explanations are fictions that mask our ignorance of the true story, which will, when we know how to tell it, have a reduced cast of characters, an exclusive set of 'purely' physical types.

This is well-trodden ground, onto which defenders of 'embodied cognitive science' have stepped. However, it is not clear who they will side with, whether they could tip the balance one way or the other, or indeed whether they will even be a unified force. A 'topographical' account of natural kinds will be developed that avoids problems other accounts face, and which is suitable for use in the general statements constructed in embodied cognitive science. Following that, we will use this account in the debate around the autonomy of special sciences in general, and the problem of mental causation in particular. The discussion will then broaden out into an investigation of causation, including a refined understanding of physical causal closure. After applying the results of these discussions to our understanding of the supervenience relation, a defensible account of emergentism will be given. Next, we will look at the kinds of properties of mental states that may be referred to in explanations of rational action, namely, the representational contents of mental states. In order to understand the nature of these states, the feedback dynamics between hierarchically structured levels of cognition involved in their evolution will be foregrounded, leading to a picture of embodied cognition that is broad and externalist. We will then look at experiential

properties, showing them to be an inseparable part of intentional states, and describing how subjecthood is emergent from brain/body/world dynamics. We will finish by outlining the implications of the refined functionalist account defended for the metaphysical notions we started with. The conclusion drawn will be that we can indeed refer to genuine mental causes which ground the non-metaphorical use of intentional explanations. Finally, I will sketch some implications for the idea of free will using empirical landmarks from cognitive science to find a path through the eroded philosophical landscape, while at the same time using these old philosophical waymarkers to guide the scientific exploration of cognition ahead of us.

Chapter 1: Natural Kinds

> Rob McKenna was a miserable bastard…. It was a particular type of rain he particularly disliked, particularly when he was driving. He had a number for it. It was rain type 17. Rob McKenna had two hundred and thirty-one different types of rain entered in his little book, and he didn't like any of them.[1] (Adams, 1984, p. 12)

Natural kinds, if they exist, are groups of objects about which we can make general statements. These generalisations form the basis of explanation and prediction, and as such, without natural kinds there can be no science. Any scientific discussion of the basis for claiming that there is such a group of things as intentional agents, as distinct from other kinds of systems that may have causal properties, must investigate the ontological basis of those properties that are taken to distinguish such systems. That is, to argue that intentional agents are real is to argue that there are certain types of states that have properties that can be called genuine causes of their effects, and which are distinct from the causal states of non-intentional objects, processes or events. Before we can state clearly what distinguishes the states of intentional from non-intentional objects, we need a metaphysical account of object classification that can accommodate both types of cases. But, before we can be clear about why we classify something as one kind of thing rather than another, we need to make sure we have a classificatory system that is defensible as a way of grouping parts of the world scientifically, and which does not beg the question one way or the other. The words we use in science to classify groups of objects, states, properties, and so on, are natural kind terms, and the class of things that fall under them, if any, are natural kinds. For something to be a cause, I will argue, is for it to be something about which true scientific generalisations can be stated, i.e. for it to be a member of a natural kind. Thus, if intentional states, such as beliefs and desires, are the causes of the behaviour of intentional agents, then such states should be natural kinds.

In this chapter I will outline an account of natural kindhood that is inclusive enough to accept the intentional kind terms of cognitive science as 'natural,' and where the true generalisations that such terms enter into are not reducible to statements of regularities containing only physical kind terms. As my starting point, I take the debate about the ontological status of the so-called 'special sciences' between Fodor (1974) and Kim (1992).

[1] The ones mentioned in the text are: 33 (light pricking drizzle which made the roads slippery), 39 (heavy spotting), 47 to 51 (vertical light drizzle through to sharply slanting light to moderate drizzle freshening), 87 and 88 (two finely distinguished varieties of vertical torrential downpour), 100 (post-downpour squalling, cold), between 192 and 213 (seastorm types), 123, 124, 126, 127 (mild and intermediate cold gusting, regular and syncopated cab-drumming), 11 (breezy droplets), and his least favourite of all, 17 (a dirty blatter that batters against one's windscreen so hard that it doesn't make much odds whether one's wipers are on or off).

Fodor's argument turns on the multiple realizability of the kinds referred to in the laws of certain sciences, using examples like oxbow lake formation and money. In both cases there are scientific disciplines that refer to these categories in generalisations, but the physical bases in which different occurrences of these kinds of things are instantiated are 'wildly heterogeneous,' in that they do not share any purely physical properties, and therefore will not fall under any purely physical generalisations. These examples could be questioned, but here I will focus on the application of this style of argument to functional kinds in cognitive science, specifically, to explanations of behaviour that refer to states like beliefs and desires. Are these explanations 'autonomous,' in the sense of not being rough restatements of complex physical explanations, by virtue of the kinds referred to being specified functionally in terms of their role within a cognitive system?

Kim's response is that since these token instances of mental kinds are physically realised, all the causal work is done at the physical level, on pain of dualism. Given a behaviour, if we had the ability to examine exhaustively all the physical causes behind this (nerve signals, muscles, neuronal activity, oxygen, food etc.), then there would be nothing left to explain; intentional explanations referring to the subjects beliefs and desires would become redundant. Eliminative materialists, e.g. Churchland (1981), hold a similar view, but Kim's argument relies on metaphysical considerations involving the concepts of supervenience, causal closure, causal exclusion, and natural kinds. In the following, I will examine each of these notions with the aim of showing that Kim's argument fails as it relies on misguided understandings of these metaphysical principles. I will show that, with a refined understanding of these, intentional explanations are non-redundant; the kinds referred to are part of the natural order about which scientific general statements can be legitimately made, and so the science that is made up of these statements is autonomous.

1.1 What are Natural Kinds?

Firstly, natural kinds are parts of the world picked out by natural kind terms. These are terms that can be used as antecedents in causal laws, parts of explanations, the bases of inductive inferences, amongst other uses. Paradigmatic examples are terms such as 'electron,' 'gold,' and 'dog.' Each of these refer to classes of things, such that if something is included in that class, then we should be able to make inferences depending on our knowledge of the generalisations that hold of that kind of thing. They can be defined as 'kinds over which numerous reliable inductive generalizations can be made' (Millikan, 1999, p. 40). Some may think that the reliance on our inductive practices in this

definition makes kinds too reliant on human understanding, rather than being fully independent of us: to be natural, categories must 'carve nature at its joints' (Plato, *Phaedrus* 265d–266a). This assumes that nature has 'joints' to discover, and if nature doesn't have clear distinctions between kinds of things, then our categories are at best intersubjective phantasies. I aim to show that including our inductive interests in the definition of natural kinds does not render them unnatural in any harmful way.

In his overview of the history of thinking on natural kinds, Hacking (1991) distinguishes four main historical approaches, associating them with Russell, Mill, Peirce and Leibniz. Peirce kinds and Leibniz kinds have essential properties that determine the manifest observational properties of members of the kind; for Peirce, these are not necessarily micro-structural features, whereas for Leibniz they are (this is essentially Putnam's position – see below). Mill kinds, in contrast, achieve membership by sharing observational properties (so both forms of jade (jadeite and nephrite) fall under the same Mill kind despite having different chemical compositions). Russell kinds form a sort of bridge between these essentialist and nominalist views by seeing members of kinds as being grouped together by their 'closeness' to other members in various ways determined by our interests, allowing microphysical properties to play an explanatory role, without begging the reductionist question (see Chapter 2: Physicalist Reductionism). Russell thought his kind of kinds could not play the role we expect of 'real,' 'eternal' natural kinds because it is an 'intensional' notion, and will change with our interests and understanding. For reasons I will explain, I think an account of this kind (which I call topographical), can succeed in giving us what we want of natural kinds, while also fitting epistemologically with the development of our (scientific) understanding of the world.

Before analysing the various approaches to natural kinds and terms, we should mention the main desiderata of any view, which is that kind terms should have explanatory value. That is, no matter how the extension of kind terms is determined (e.g. by similarity to paradigmatic examples or fitting theoretical definitions), or how candidate kind terms are selected (determined by our interests or facts about the natural world), whether we are talking about grouping objects, or events, or properties, or indeed whether there are degrees of naturalness (LaPorte, 2004), natural kind terms must be useful as a basis for inductions. That is, they must be 'projectable': determining that a particular thing falls under a kind label should allow us to make predictions about unobserved properties of that thing, and unobserved instances of that kind, in other words, to make generalisations about things included in that category.

1.2 Are Natural Kinds Found or Made?

Two main positions on the subject of natural kinds have been developed historically; I aim to show that there are also ways to steer a course between them. On the one hand, there are realists, starting with Aristotle (Ayers, 1981), who think that natural kind terms 'cut nature at the joints,' i.e. refer to objective ontological categories, generally characterised by shared essences. For them, natural kind terms pick out ontological entities which have their form essentially. Non-essential properties of things are 'accidents,' and any definition in terms of these is called nominal rather than real.

In opposition are non-realists (or nominalists) who believe that there is no fundamental distinction between real and nominal definitions: natural kind terms refer to subjective, epistemological and observational categories. On this view nature has no joints; where the lines are drawn between kinds is down to us, based on similarity relations in experience. In reality, differences are always a matter of degree rather than kind; everything is 'on the spectrum': 'There are Animals so near of kind both to Birds and Beasts, that they are in the Middle between both' (Locke, 1700, III, vi, 12).

My position is between these extremes. The world may not have neat boundaries between kinds of things, and so we may contingently define such boundaries to enable us to navigate the world, to give directions to each other, but once these boundaries have been drawn, it is an objective fact, that depends on the actual structure of the world, what falls on one side or the other. So far, the nominalist can accept this. The difference is that my position is realist about there being an explanation of how the differences have developed naturally. However, it is not a straightforward realism, as there is no fact of the matter about which classificatory scheme is objectively preferable; this depends on us, our physical embodiment and cultural embeddedness. There is not one ideal scheme towards which our developing science is incrementally moving.

This is the topographical account of natural kinds: like a map of a territory, we decide where one feature ends and another begins, depending on our navigational interests, but not arbitrarily, as these decisions are based on the real features of the landscape. Moreover, our 'map' is refined over time as people use it to navigate; previously unnoticed pitfalls are drawn in, misplaced peaks redrawn, outcrops named. Before further defining this position, I will add some detail to the picture of the rival positions.

Frege's descriptivist theory can be seen as a refinement of the nominalist view. Words have 'senses' that exist intersubjectively, and these determine the reference of words in virtue of objects satisfying the descriptions attached to these senses. Kind terms express concepts of the classical

sort (Margolis & Laurence, 1999): they are associated with definite descriptions and these supply the wielder of that concept with the means to determine whether any given object falls within the extension of that term based on manifest, observational properties. (I am ignoring the distinction between kind terms and proper names, and assuming that the names for classes function like the names for individuals; these classes being made up of all and only those individuals that satisfy the definition of that class.) For example, the reference of the kind term 'tiger' is fixed by descriptions like, 'large, carnivorous, striped cat,' and we learn how to use the word through learning the descriptions associated with it. This has the advantage of enabling us to refer to things that we have had no direct experience of: we can talk about Pluto, because we know that it is the ninth planet from the sun. Furthermore, the fact that we use the 'sense' (i.e. meaning in terms of definite descriptions) to fix the reference of a word, rather than through some sort of direct connection between word and object, explains why we are not automatically aware if two words, with different definitions, happen to refer to the same object. The classic case of this is Hesperus and Phosphorus, the former defined as the last star in the evening sky looking west, the latter as the first star in the morning looking east, but in fact both of them referring to Venus, which was only discovered after millennia of using those words to successfully refer to that object in the sky. Finally, this enables us to talk about non-existent things, like unicorns, without being accused of making meaningless utterances; the words have sense, but no reference, as no such things satisfy the descriptions associated with that word.

However, there are several problems with this kind of theory. Firstly, it is not easy to specify these conditions for inclusion/exclusion as it seems it is always possible to find exceptions to supposed necessary conditions (e.g. not all tigers have stripes), and we are able to deploy terms without knowing the descriptions supposedly attached to them. For example, someone may say, 'Brighton contains the last remaining population of native elm trees in England' without being able to describe or point out elm trees. Neither is an object's satisfying the sufficient conditions for being a particular kind of thing actually enough. For example, in the case of an alien 'dog': it may bark like a dog, but that does not mean it is a dog.

There is also the related problem of vagueness. Even if we could define them, sets don't easily accommodate vagueness, making application to the messy real world, where there are few neat boundaries, problematic. Most biological kinds, being the product of the gradual process of evolution, could be ruled out: the ancestor of the tiger didn't suddenly go from satisfying one set of descriptions to another. Furthermore, approaches that rely on definitions cannot account for stability of reference through theory change. If our definitions change, we would, by definition, not

be talking about the same things anymore. This kind of incommensurability is undesirable, since we do think that we are talking about the same things as earlier people were when they referred to whales, even though we no longer think of them as fish.

This is the essence of Kripke's modal argument: it is possible for something to be a member of a kind without fitting the description, and possible for something to fit a description without being a member, because names (including kind terms) are 'rigid designators' (they refer to the same thing in all possible worlds), whereas descriptions are not rigid in this sense. Put another way, names don't connote, they denote. Natural kinds are such that science discovers about them necessary but non *a priori* things (e.g. the chemical composition of water). So, to use Putnam's (1975) example, 'water' on Earth refers to H_2O, which satisfies certain descriptions (e.g. wet stuff that falls from the sky); 'water' will continue to refer to H_2O whether or not is satisfies the descriptions we associate with it on the surface of earth, but it is possible that some other stuff which is not H_2O (and therefore not water) satisfies those descriptions elsewhere (so those descriptions will pick out some other substance in that context).

Kripke and Putnam's alternative to descriptivism is the causal theory of reference, which claims that natural kind terms refer to all things that share essential properties with the object that the term was first coined to refer to, no matter what 'sense' speakers attach to the word. Kind terms are causally connected to an act of ostensive 'baptism' (someone points at something and decrees that all such things shall be known as X); objects are of the same kind if they share a micro-physical essence with the object originally so christened. These theoretically interesting essences are discovered by empirical investigation after this act, rather than being stipulated by definition. What is being referred to depends on the speaker's intention as to what they are talking about, but does not depend on the definition that speaker has in mind. (This is to assume, for the moment, that a referential intention and the knowledge of the intender can come apart unproblematically.) So neither the originator of the usage, nor those who learn to so use the word from her (directly or indirectly), may be aware of what this essence is. The relationship between names and essences is one of necessity due to their being 'rigid designators': nothing could possibly be water and not be H_2O, and *vice versa*. The advantage of this is that there will be no change of meaning with a change of our theories and therefore of our understanding, and no associated problem of incommensurability.

A further benefit of a causal theory of reference is that it gives objective criteria for including things in sets in terms of physical properties that allow us to get past inductions from sensory properties, to a situation where the terms of mature science pick out things with dispositions for which we can

state the mechanism by virtue of which they are disposed to be such-and-such, and the dubious notion of descriptive similarity is disposed with. This allows us to rule out accidental regularities, or temporary, contextual effects (Millikan, 1999, p. 52). It also grants us the ability to talk about things we know nothing of, but still say something with meaning. This is because we can rely on the fact that somewhere back in the line of causation of the use of the term there stands someone who knew what she was talking about (in terms of knowing the reference of the term, if not the essential properties that all objects of that kind share).

Similar claims could be made for theoretical terms, such as 'gravitational waves,' the 'essence' of which is defined theoretically, but which is not understood by many users of the term. This seems to be a kind of descriptivism, where the description is understood by a few experts who coin it and continue to use it. Such cases rely on the tokening of mental states in a group of experts, rather than being a simple label with a usage that is passed along the line. There are complicated issues here that I will put aside for the moment.

Although causal theories of reference seem to avoid the problem of our being able to deploy kind terms without needing to have the associated descriptions in mind, the understanding of the users of such terms cannot be totally ignored. Back down the causal chain, when the coiner of the term pointed towards a portion of the world and made a sound, how can we know what he was pointing at unless we ask him to describe it? This is the *qua* problem (Devitt, 1981), or the problem of ostensive definition (Wittgenstein, 1953, §§ 28-30): as *what* is that part of the world being picked out? Is he pointing at the colour, the leaf shape, or is he intending to name that species of tree? Similarly, there is the 'which' problem: which thing are you actually pointing at, the tree or the wood? In other words, under which sortal is the ostension being made? It is unclear how these questions can be answered without referring to the intentions, and therefore the beliefs, of the pointer. Also, incommensurability threatens again given subsequent changes in belief systems and the uses of words: if the extension of a term depends partly on the theory-laden descriptions connected with it, then referrers with different theories may fail to communicate successfully when they use that term. It seems that Kripke and Putnam are guilty of assuming that non-linguistic acts like pointing are not 'contaminated' by the linguistically acquired beliefs of the pointer. Kuhn, who introduced the term 'incommensurability' to the philosophy of science, came to realise this:

> My original discussion described non-linguistic as well as linguistic forms of incommensurability. That I now take to have been an overextension resulting from my failure to recognise how large a part of the apparently non-linguistic component was acquired with language during the learning process. (Kuhn, 2000, p. 315)

The kind of learning process mentioned was evident in his earlier work, in the use of exemplars in educating scientists (Kuhn, 1962); through learning canonical applications of a theory, the norms that determine the use of terms are learnt.

Even if it were possible to unproblematically pick out kinds ostensively, this assumes that there is a micro-physical essence to all natural kinds worthy of the name. Is micro-physical essentialism something we can assume *a priori*? Not if we don't want to beg the question against certain forms of non-reductive physicalism. In the case of biological kinds, the candidate essences would, I assume, be DNA. As we will see in the next section, this is an oversimplification, but for now, it is clear that the metaphysical arguments for a causal theory of natural kind term reference are problematic. It seems that causal theories of reference cannot be 'pure': in order to fix reference we need to know something about the descriptive information connected to terms. Put another way, to understand the speaker's intentions, we need some access to the speaker's intensions. Moreover, we will see below that the causal theory also seems to be at odds with the way scientists actually operate.

1.3 Rigidity

Rather than the rigidity of designation described, we find that science refines its beliefs about archetypes upon discovering new specimens that are causally 'downwind' of the original coinage (Mellor, 1977, pp. 301-4) (see §1.5.3 Pluto: A case study in natural kind term demarcation). As mentioned, some natural kinds have no ostensive archetypes, e.g. neutrinos when they were merely theoretical entities. Also, no reason is given why certain properties must be shared by all the members of a kind, rather than members being related by 'family resemblance' (Wittgenstein, 1953, §§65-71; Mellor, 1977, pp. 305-6). Mellor concludes: 'The stock candidates for essential properties, are either not even shared in this world by all things of the kind, or their status is evidently more a feature of our theories than of the world itself' (Mellor, 1977, p. 311).

Despite the ability to cope with theory change being advertised as one of the strengths of the theory ('Causal theory is supposed to assure *continuity* of meaning and reference.' (LaPorte, 2004, p. 112)), there is the problem that the causal theory has difficulty with cases where terms are refined due to theory change, resulting in the need to decide which objects will continue to be included in the original category, and which excluded, as happened with debate of the planetary status of Pluto (see §1.5.3 Pluto: A case study in natural kind term demarcation).

Kind terms may be 'rigid designators,' but not independently of fixing some definitions, which might depend on facts about us, the definers, as well as the things we define. Once fixed, it may be clear what a term does and does not refer to in all possible worlds 'nearby' enough to ours to count. Thus, rigidity is not what distinguishes natural from nominal kinds, rather this is achieved via a hybrid theory of reference that takes into account the intentions of the original ostensive definer and not just the microphysical essence of what was pointed at (this will become important later in the discussion of Kim's arguments in §2.1.1 The 'Special' Debate). Thus, causal theory doesn't get rid of theory-laden descriptions, and reference is not determined just by the way the world is. The nature of the original paradigmatic samples may partly determine meaning and reference, but this doesn't block referential change:

> Causal baptisms, which according to the causal theory endow terms with their reference conditions, are performed by speakers whose conceptual development is not yet sophisticated enough to allow the speakers to coin a term in such a way as to preclude the possibility of open texture, or vague application not yet recognised….Speakers are left to decide whether to call the monotremes "mammals" on the basis of whether they have the right features, which… seems to leave the matter open….Progress is… replacing vague statements… with straightforwardly true statements…. *Whichever*… refinements might have been adopted….The decision to use a term to designate one thing rather than another cannot be false. It is the statements made with a term on a use that can be false. (LaPorte, 2004, p. 134)

So far I have glossed over the distinction between names of individuals and names of categories, but there are important differences. In the case of the rigidity of singular terms like proper nouns, it is clear what remains the same through different uses of the word, but it has been claimed that it is not clear what remains the same in the case of kind terms (Schwartz, 2002). A term can be said to be rigid if the reference remains the same in all possible worlds (or all relevantly nearby possible worlds). The problem, it is claimed, lies in distinguishing between general terms like 'bachelor,' which are purely nominal, and general terms like 'tiger,' which are supposed to be natural, that is, not dependent upon our definitions: broadening the notion of rigidity to all general terms, including ones dependent solely on conventional definition, is to trivialise it. The point of rigidity, according to this argument, is for it to do the work of differentiating between natural and nominal kinds. The extensions of general terms that are determined by our definitions change as definitions change, but the extensions of natural kinds are determined by the structure of the world *and* a naming event. LaPorte (2000, pp. 297-99), in contrast, does not see this broadening as trivialising the notion of rigidity, as some general descriptions still come out as non-rigid. So, while 'honeybee' is rigid, the co-extensive 'the insect farmed for honey' is not, as other insects may play this role in other possible worlds.

Generally speaking, it seems true that all kind terms have an aspect of the perspectival, depending on our definitions, as well as some rigidity, depending on the structure of the world. In the case of

the term ‘planet,’ for example, the definition may be a matter of decision, but once it is fixed, it is clear what the term refers to in all possible worlds. The same is true of ‘bachelor’; its meaning is a matter of convention, but it can be said to have some rigidity, since in all possible worlds where there are males, and a practice like marriage exists (that is, in all nearby possible worlds), it will refer to all and only those males who are not married.

This looks like a sort of two-dimensional theory of meaning (Lewis, 1966), in which terms are taken to have a primary and secondary intension, where the primary intension is that internal, reference fixing criteria that the user of the term has *a priori* access to as a competent member of a linguistic community, and the secondary intension is the ‘modal profile’ of the term, that is, what it takes to be a member of that kind in all possible worlds, which speakers have *a posteriori* access to. There are two main varieties of two-dimensionalism, namely empiricist and rationalist varieties. I will restrict myself to the empiricist variety here since our ultimate aim is to apply this theory of kinds to explanations of human actions, and it is the empiricist variety that foregrounds the causal factors involved in individual usage. For our purposes, I will take the criteria of application of a kind term to be a potentially implicit set of pluralistic associations (definitions, prototypes etc.), embedded in general usage and the wider communicative practices and empirical concerns of society. So, the primary intension may be a set of implicit conventions which can be subjected to conceptual analysis by the user, and could change given evidence. However, such *a posteriori* change doesn’t necessarily bring us closer to the original meaning of a term, instead changing that meaning.

Taking the example of water, the primary intension that fixes the extension of that term for speakers may be something about potability, falling from the sky, translucence, and so on, and we learn to apply that term in the process of learning to use the English language. Now, it may also be the case that our society has, since the scientific revolution, decided that some of our words should have their extension determined by being members of a chemical kind, and the word ‘water’ should henceforth refer to all and only that stuff that shares a chemical constitution with the stuff that in our world is potable and falls from the sky, i.e. H_20. The term ‘water’ has then been rigidified. But it wasn’t necessarily so. In the case of jade, its usage has not been restricted to a chemical kind. So, in the case of natural kind terms like ‘water,’ the fact that it is an *a posteriori* truth that ‘water = H_20’ is not necessary; it is a contingent fact that for scientific purposes we have rigidified the term in that way. Moreover, it is an open question as to what actual substances in the world we will allow to be included in the extension of the term, given that there are isotopes, etc. (see §1.5 Mapping). The rigidity of the secondary intension is contingent on our practices just as the criteria of application of the primary intension are.

It is the mapping out these various criteria in a multi-dimensional conceptual space that gives us the topographical picture of kindhood. Restricting ourselves to three dimensions for the moment for the purposes of ease of exposition, the three axes could be as follows:

1. The x-axis represents similarity of primary intension. For example, how similar is it, in terms of perceptual properties, to prototypical water?
2. The y-axis represents some kind of 'modal distance.' For example, are there possible worlds in which the stuff in question falls under the rigidified term 'water'? XYZ would be ruled out despite looking and behaving like water; an isotope that is radically dissimilar (e.g. pink and fluffy (LaPorte, 2004)) would be ruled out despite being chemically close; but an isotope that is liquid and see-through may not be.
3. The z-axis represents projectability, or inductive utility. On the summit of the lump in intensional space that represents 'water,' are the clear cases of water found on earth that are composed of just H_20, are liquid, translucent, and thirst quenching, and which will confirm all the generalisations we make about water.

The advertised benefit of rigidity was to capture the necessity of identities discovered *a posteriori,* like that between the two names, Hesperus and Phosphorus, both of which refer to Venus (Kripke, 1972). A similar case could be made in the case of the names for categories of things that we have decided should be defined by their sharing a physical or chemical nature. We have already seen that even in these cases the straightforward semantic externalist account is an over-simplification. The next question is how our accounts of the meanings of kind terms apply to other, less basic kinds, like biological categories. According to Schwartz (2002, pp. 270-271), an identity like 'the honeybee = *Apis mellifera*' is not like 'water = H_2O,' in that the Latin name is just a name, rather than some biological microphysical property. He assumes that there must be an equivalent identity between biological species and microphysical types, namely DNA. LaPorte (2000) instead talks of the cladistic essence of biological kinds that involves belonging to a lineage, rather than merely being another name for the same thing, in which case the Latin name does contain information that could turn out to be false, i.e. the assumed lineage. But rigidity cannot do the work of distinguishing natural kinds from nominal ones (LaPorte, 2000, p. 304), for reasons given above (in 1.3 Rigidity), this being left to a non-essentialist causal theory of reference, where a term is a natural kind term because we have decided to use it as such in our linguistic community.

In the case of biological kinds, baptism requires some descriptive information: e.g. 'That is species X' describes the referent as a species. Such natural kinds are not eternal and immutable; biological kinds are paradigmatic natural kinds, with, if lineage is important to classification, historical

essences. Moreover, we do not discover facts about these essential properties *a posteriori*: 'Mammalia = the clade that stems from ancestral group G' is not discovered to be true, rather, 'these terms have undergone meaning refinement to make them refer to the relevant clades…. They are rigid *de jure*, rather than rigid *de facto*…. A term need not keep its meaning over time in order to be rigid at a time' (LaPorte, 2004, pp. 48-49). When new evidence turns up there is a choice to narrow or widen the term; we have to ask why it belongs to that particular taxon: 'If neither the biological species concept, or the phylogenetic species concept can be discovered to be true, and each stipulate a different essence for the species, how can it be said that scientists discover the essences of previously baptised species?' (LaPorte, 2004, p. 73). There are competing definitions of species, e.g. clades vs. grades, where a clade is an ancestral group and all of its descendants and a grade is a group with a shared level of organization. One system may come to be accepted, but this will not be because one is necessarily true in that it is the correct representation of the way the world is organised, but rather due to such qualities as its ease of use and 'tidiness.' In this way, 'the ranking of groups [may be] largely arbitrary… [but] this casts no doubt on the naturalness of ranked groups' (LaPorte, 2004, p. 186 note 19). Moreover, given that the proposed ranking may be discovered to be too vague to be useful, in order to dispel this vagueness we may need to modify the meaning of the term, but this refining cannot be said to be approaching more closely the originally intended meaning of the term (LaPorte, 2004, p. 90).

1.4 Projection

The reason natural kinds matter is because we can make generalisations about them, meaning we can explain what happened, predict what will happen in future cases, and, if we desire, intervene in the course of events. In other words, natural kinds are what science is about. Terms that can be used in this way are said to be 'projectable,' and the process of making general statements from observations is inductive inference. Therefore, before continuing we need to deal with the problems of induction, old and new.

The problem of induction, as raised by Hume (1738), asks how we can justify going from finite observations to a fully general conclusion, as we can never know for sure that the next observation will not turn out different from previous cases. Umpteen white swans may be observed, but we can never validly conclude that all swans are white, so scientific generalisations cannot be facts that we know.

One solution, confirmationism (Carnap, 1950), is to take instances of a law as confirming (raising the probability of the truth of) that law. So the hypothesis 'All emeralds are green' is confirmed by finding more green emeralds. At no point can we definitively state that such generalisations are known facts, but that doesn't matter as long as we are satisfied with well-confirmed generalisations that we can use in scientific practice.

But why do we think some generalisations are confirmed by their instances and not others? This is the new riddle of induction (Goodman, 1955). Not all statements containing projectable predicates are confirmed by their instances, that is, they are not all 'lawlike.' For example, 'Green things are non-intelligent' is accidental, and observations of non-intelligent cucumbers do not confirm it. How can we tell that 'Emeralds are green,' on the other hand, is lawlike? What about the hypothesis 'all emeralds are grue,' where 'grue' 'applies to all things examined before *t*, just in case they are green but to other things just in case they are blue' (Goodman, 1955, p. 74). So, instances of green emeralds examined before *t* will confirm this hypothesis too. This means that any particular observation could be seen as confirming any number of hypotheses, and there is no reason to see such observations as lending credence to any one of them in particular.

The idea of 'projectability' is central to the solution to the new riddle of induction. The difference between 'green' and 'grue' is that 'green' is a projectable predicate, in that it is one on the basis of which we can make correct inductive inferences. That is, on the basis of observed tokens of objects with the property of being green, we can say that unobserved tokens of the same kinds of objects are likely to be green too. It is information that we can use to plan ahead. 'Grue' is not projectable as the observation of a green token instance of a kind of thing made before an arbitrary time in the future does not support the prediction that an observation of another token of that kind made after that time will be blue.

Quine puts it thus: '…projectable predicates are predicates ζ and η whose shared instances all do count, for whatever reason, toward confirmation of [All ζ are η]' (Quine, 1969, p. 115). For example, 'All ravens are black': we can expect that if we see a raven it will be black, that if we see a non-black thing it will not be a raven; but we cannot expect non-ravens to be non-black, or black things all to be ravens. In other words, projectable predicates are those that are usable in valid inductive inferences from finite observation to general statements.

According to some, e.g. Israel (2004), 'grue' has often been misinterpreted (e.g. by Kripke) to mean 'X is green prior to time *t*' without entailing anything about being examined. This definition makes 'grue' a two-place predicate, its extension being the set of ordered pairs <X,T> such that X is green

and T < *t* or X is blue and T ≥ *t*. Generalisations involving this predicate entail that every emerald will change its colour from green to blue at time *t*. Israel claims that to be relevant to the wider problem of induction, not just prediction, the hypothesis 'all emeralds are grue' needs to be accidental, unlike the lawlike 'All emeralds are green.' If we accept the interpretation of grue where emeralds suddenly change colour at *t*, then this misses the fact that the property of being observed before *t* is crucial, making it a problem about prediction, rather than about generalising from observed to unobserved instances (Israel, 2004, p. 335). Framing it the wrong way weakens the riddle by requiring objects of a certain sort to change their colour for no reason. To say that all objects of a certain kind will suddenly, and accidentally, change their colour at the same time, is 'strange,' and weakens the force of the argument, according to Israel.

Does it really matter that it's strange? It's also strange if we 'just happen' only to have found green emeralds so far, on the assumption that there are blue ones out there too that we have, by accident alone, not seen. There may be an explanation for why this is so (in which case it is not purely accidental, as Israel wants), and if it is just accidental, then that makes it a sampling problem rather than a deeper one about justified generalisation. But equally, there *could* be an explanation for colour changing at a certain time. The problem with Israel's way of framing the new problem, is that it's not really new if put that way. The fact that some emeralds may be blue, although they haven't been observed yet, is exactly analogous to the situation with swans before the discovery of the black variety. Moreover, since there is only an accidental relationship between the colour and time *t*, there is nothing special about the colours being proposed; there is no reason to call this property 'grue' rather than saying some emeralds are green, others are blue (after all, we don't feel the need to say that the colour of swans is 'whack').

But, putting aside such concerns for the moment, Israel wants to say that the definition of grue implies there are two kinds of emeralds, green and blue, and that by accident we have found only green ones and will continue to find only green until *t*. Since an accidental generalization like this is not confirmed by its instances, why do we believe some hypotheses are lawlike (i.e. confirmed by their instances)? The problem is that we can make contradictory predictions from the same evidence if we accept confirmationism (the view that the likelihood of a given hypothesis being true increases with each observation that confirms that hypothesis). Only predictions based on lawlike propositions are valid, but we don't have an independent criterion for lawlikeness. As mentioned, Goodman's solution is to distinguish between projectable & non-projectable predicates. If 'All emeralds are green' is lawlike, then 'green' is projectable. But 'All green things are non-intelligent' is not lawlike, and the fact that all observed green things have been non-intelligent is accidental. So,

either ‘green’ is not projectable after all (i.e. its instances don't confirm generalizations it enters), or projectability of the predicates in a universal proposition is not a sufficient condition for lawlikeness.

Davidson (1966) says projectability is not a property of a single predicate, but a relation between predicates: ‘green’ is projectable for emeralds, but not for intelligent beings. But here is an impasse: if projectability is a property of a single predicate, we cannot use it as a criterion for lawlikeness, but if projectability is a relation between predicates, it becomes uninformative (since lawlike propositions are those with projectable predicates, and projectable predicates are those found in lawlike propositions).

Israel’s solution is to take projectability to apply to whole propositions rather than predicates (Israel, 2006). We project ‘all emeralds are green’ because we believe the proposition is true and lawlike. But we still face the problem of distinguishing lawlike from accidental propositions. Lawlikeness is defined by how something would fit into our ‘whole epistemic web,’ with inductive practices linked to the (practical) possibility of explanation: ‘A proposition is lawlike if it is rationally explainable (or would be if it were true), but the property of being rationally explainable is relative to the context’ (Israel, 2006, p. 276). We ‘project’ generalizations rather than properties: sometimes correlations prompt us to look for explanations (e.g. smoking & cancer), and may lead us to revise our epistemic web; sometimes they don’t, or shouldn’t. Projectability ‘depends on context because it depends on the entire ‘web of belief,’ including the epistemic values of reflective equilibrium, coherence, simplicity...’ (Israel, 2006, p. 276).

> ‘Strange’ generalizations cannot be preferred because of the ‘cost’ of including them in our web of belief. The practical possibility of explanation means that relative to a web of belief B, it is practically possible to explain the truth of generalization A, iff B implies that (1) A is in fact true or has a high likelihood of being true; and (2) B implies that it is possible to explain the truth of A on the basis of facts... that actually obtain. (Israel, 2006, p. 279)

He concludes that not every generalization is confirmed by its instances, only lawlike ones, that projectability is not constitutive for lawlikeness, rather lawlikeness determines projectability (lawlikeness being determined by the explainability of the truth of a generalization) and that ‘an acceptable system of inductive logic that takes only the logical forms of the evidence and the hypothesis into account is impossible’ (Israel, 2006, p. 283). (We will return to the subject of explaining in §3.2 Explanation.)

If this is right, projectability is a matter of explainability within our worldview. One of the central beliefs we have is the principle of sufficient reason: that events happen for reasons. So, if emeralds are grue, there must be a reason, which we should be able to cite, why they will change colour, or why we have only observed green ones so far. In either case, there must be some ‘local’ reason that

we can, in principle, observe. It must be spatially local, otherwise it is not a property of emeralds but of something else, which means that it must also be temporally local (i.e. present in emeralds before *t*). In order to justify a 'gruesome' claim, then, we would have to be able to point towards a possible present property to explain the properties that go along with being grue.

Put another way, 'an observational predicate is apt for induction to the extent to which one can locally determine whether or not the predicate holds in a given case' (R. Chrisley, 2008, pers.corr., 1 Apr.). This is to take an experientialist view of observational concepts, in which, in order for one to be said to possess a concept, one needs to be able to say one is experiencing an instance of that property on the basis of that experience alone. A full defence of such a position is beyond the scope of the present work, but has recently been made in a thesis by Ivan Ivanov (2016, p. 209): 'Having an observational concept… involves having the experientially-based capacity to determine, without further empirical investigation, what the property picked out by the concept is.' He gives this summary of his position:

> I. Having an observational concept involves having substantial knowledge of the property picked out.
>
> II. Such knowledge consists in the capacity to come to know what the property is without further empirical investigation.
>
> III. Experience of instances of the observational property in optimal viewing conditions plays an essential role in providing the basis for the capacity to come to know what the property is without further empirical investigation. (Ivanov, 2016, p. 204)

If we accept such a view of observational concepts, then we only need look (in optimal conditions) to see if an object is green, but before saying an object is grue, we have to know the time and whether it was previously observed. When a difference like this is noticed between 'normal' predicates and 'gruesome' ones, a standard response is to reassert the riddle by redefining the 'normal' in a 'gruesome' way: if we claim that grue is disjunctive (in that it is defined in terms of being green *or* blue) whereas green is not, it can be replied that it is possible to define green in terms of being either grue or bleen ('grue and observed before *t* otherwise bleen'). That reply may work against an argument that claims that the main problem with 'grue' is that it is disjunctive, but not against one that sees the main problem as being that we should be able to locally determine if an observational predicate holds or not. The advantage with the local solution is that redefining in this way does not make green non-local, so we can take as a starting point that being locally determinable is a criterion for a predicate's being projectable.

It might be asked, though, whether there are some predicates of kind membership (e.g. 'being a dog') that are not locally 'observational,' but rather, historical. Something that looks and smells like a dog, and is observationally identical down to what can be observed about its micro-structural

composition, is not a dog unless it shares a common ancestor with other dogs. Is an artificial dog that is exactly similar to a real dog, with respect to locally determinable properties, a dog? Predicates like these could be called 'locational' in Swinburne's (1968) distinction between these and what he calls 'qualitative' predicates, and concludes that where there is a conflict, we should prefer qualitative over locational for the purposes of projection. Swinburne uses 'qualitative' in a broad sense to include any kind of possible observations, not merely 'naked eye' ones, so the correctness of application of a qualitative term can be determined without knowing anything specific about the time or place of the occurrence, and can include, potentially, dispositional properties, like 'being brittle' (which could be tested by striking). Interestingly, this notion of qualitative is relative to the development of our instruments of observation, and so the wider network of scientific beliefs.

To the objection that membership of an evolutionary kind is not locally determinable, then, we can reply that although our explanatory practices refer to the evolutionary past, how we determine whether a creature is a member of a particular evolutionary lineage is generally by examining facts that are locally determinable, e.g. about its DNA, as well as what it and its parents look like. These count as locally determinable pieces of evidence for its belonging to that 'historical' kind. The fact that it is conceivable that an exact copy could be made just means that we could be fooled, rather than being a reason for rejecting this notion of kindhood altogether.

Another way we could approach the problem is to allow a distinction between 'pure' and 'impure' properties, where the pure properties are non-relational, and all other properties are impure. This latter kind, though, may still be contained in the broad definition of observational outlined previously.

Besides, whether something has the appropriate lineage probably won't make a difference when it comes to projectability. If it's physically identical to a real dog, it will bark in appropriate circumstances, and it won't become a cat at midnight. Or if it were to, then we would expect to be able to find some properties present in the creature that would explain this and enable us to, in principle, predict that. There would be some physical mechanism keeping track of the time for example, or one that is triggered by external events coupled with that time (e.g. the striking of the clock 12 times). The general point is, kinds are generalisations made by us to fit in with our theoretical understanding of how the world works, which includes a general physicalism that states that causal properties must be physically realised, which entails that we must be able to locally observe the relevant causal factors that enable us to say that an object belongs to such and such a kind, on the assumption that we have the necessary observational apparatus. Such historical

properties are either of interest etiologically, or in allowing inference to some property shared with creatures of that kind, justifying projection.

To recap: we are looking for a way to ground the projectability of properties as the epistemic criterion for natural kindhood in such a way that it can be used to make valid inductions from one token of a kind to the next, and the reason we are looking at natural kindhood is to allow a detailed analysis of a form of argumentation found in the reductionism/non-reductionism debate that focus on the relationship between lawlike statements made at different levels. The expectation that we will encounter the same properties when we next encounter an object of the same kind depends on the idea of observational resemblance. According to one of the main proponents for reduction, the ontological criterion for this resemblance is the sharing of a causal nature, based on the principle of 'same cause, same effect': 'Causal powers involve laws, and laws are regularities that are projectable' (Kim, 1992, p. 525).

The other main antagonist in the debate about the metaphysical status of 'higher-level' laws, or the autonomy of 'special scientific' laws, Fodor, defines natural kind terms as 'projectable' terms that enter lawlike statements in a scientific discourse as antecedents or consequents (Fodor, 1997). Assuming that underlying lawlike regularities referred to in scientific discourse are causal relationships, this is one point that he and Kim can agree on: the principle of causal individuation of kinds: 'Kinds in science are individuated on the basis of causal powers; that is, objects and events fall under a kind, or share in a property, insofar as they have similar causal powers' (Kim, 1992, p. 522). (What they don't agree on will be explored in detail in §2.1.1 The 'Special' Debate.)

Science carves nature up into kinds using empirical methods based on epistemic criteria, like predictive and explanatory power, and ontological beliefs. It explains how the things we find in the world came to be and be as they are, and what the effects are of their being thus (for understanding and control). Hume (1738, Book I, Part III) called causation the cement of the universe, but his regularity account of causation leaves us unable to explain singular causal statements or partial regularities (see §3.1 Causation), for which the notion of kinds is 'the link between singular and general causal statements' (Quine, 1969, p. 239).

As mentioned, Fodor (1974) explains kindhood in terms of laws: a given predicate P is a 'kind predicate' of a science just in case the science contains a law with P as its antecedent or consequent (which is in line with a generally Quinean ontology). For an autonomous science, the laws containing such kinds should not be replaceable without loss of predictive power by more 'basic' statements containing the kind terms of another, more fundamental description: causal properties are

properties of objects which persist in their own right, not mere aggregates of the causal properties of their constituents. (These issues will be explored in Chapter 2: Physicalist Reductionism and Chapter 3: Levels of Causal Explanation.)

According to this position, we should restrict ourselves to talking about causal properties to avoid 'subjectivising' the notion of kindhood. Things are similar in a cosmically primary sense, then, 'to the degree that they are interchangeable parts of the cosmic machine revealed by science' (Quine, 1969, p. 240). (I will put aside for the moment the quibble that being 'interchangeable' could be interpreted subjectively, and that causal statements depend, to some extent, on the taking of a perspective.) Given that the laws of science are causal, and that the existence of laws requires 'independent fundamental magnitudes' (Putnam, 1969, pp. 249-51), to replace each other (to be tokens of the same type) things must share a causal nature. But to avoid deflating the notion of kindhood through being too strict in the interpretation of 'sharing a causal nature,' we have to say that they have *similar* causal properties, which re-introduces subjectivity, as what counts as similar enough depends on judgements we make, according to our own explanatory interests.

The reason we expect the next emerald to be green rather than grue is based on causal similarity, but it is, according to Quine (1969) this that creates the puzzle, as it seems there is no respectably scientific way to cash out the notion of similarity: how do we draw the line between things close enough to the paradigm and far enough from the 'foil'? How do we choose a paradigm? We start with our innately endowed, perceptual similarity space ('good for food-gathering'), then we start to theorise, and adopt new groupings on the basis of those that are inductively fruitful, and so those become 'entrenched' (in Goodman's terminology). Statements like 'emeralds are grue' are not entrenched in accumulated scientific and linguistic practice. However, according to Quine, similarity should not be defined relative to our theorising, but must be a fact about the objects and their properties: '[Similarity] belongs in the subject matter not of our theory of theorizing about the world, but our theory of the world itself. Such would be the acceptable and reputable sort of similarity concept, if it could be defined.... It does get defined in bits: bits suited to special branches of science' (Quine, 1969, p. 240). Quine wants to get beyond inductions from sensory properties, to where the terms of mature science pick out things with certain dispositions for which we can state the mechanism by virtue of which they are disposed to be such-and-such.

This, I will claim, can be achieved by the Virtual Machine Functionalism I will argue for later (§6.1 Virtual Machines), without having to reduce everything to physics or throw out the notion of similarity all together. On the topic of functional kinds, many, including Fodor (1994, p. 31), would distinguish them from natural kinds, the latter being defined by similarities in their microstructure,

as opposed to their causal dispositions. This is, in my opinion, giving up too much ground to the reductionists (strange for Fodor to do so), as I don't think a clear distinction can be drawn. I will turn now to describing my own position that, I claim, manages to find a way between the problems of the above account.

1.5 Mapping

The account presented and defended in the following will describe our space of natural kind concepts as being analogous to a map: something created by us for our use in navigating this world, the nature of which depends on both our needs and the structure of the world. I will call accounts of this type topographical.

A topographical account is not realist about the existence of kinds as abstract objects (universals in old money): kinds in that sense don't exist separately from our construction of them. But it is realist in the sense that the tokens so classified are such that they objectively have certain properties that mean they fall under kind terms once we have fixed the meanings of those terms. Neither is this account nominalist, even though nominalists could accept the points just made, as the categories are not constructed by us before they get applied to the world; they emerge as the result of a reciprocal, feedback dynamic through evolutionary, socio-cultural, and developmental timeframes. This allows the world to play a formative role in the construction of our world-picture, and consequently a normative role too. Nor is it conceptualist, despite agreeing that our perception of the world is penetrated by our understanding, this penetration is not 'all the way down,' but rather a negotiation whereby the map may draw our attention to certain features of our experience, but it is not the only means by which we engage with the world, and some of our experience may be unconceptualised despite being somewhere on the map (a place with no name).

Natural kinds don't necessarily get at nature's joints, because sometimes there aren't any joints but rather a collection of individuals along multiple continuums of similarity. Some of these individuals are common enough, similar enough, and important enough, for us to feel the need to classify them as kinds. The problem with the traditional conception of natural kind terms is the requirement to state necessary and sufficient conditions for the application of the term. The world is uncooperative, refusing to fit neat categories that the discrete nature of language demands. Our kind terms should rather be seen as picking out 'geographical' features in a topographical space, like naming mountains, and parts of mountains in a landscape: the tops 'stand out' as features, but at some, ill-defined point the slopes become the surroundings, the boundary being drawn by us. This

may seem unsatisfactory, but fits with experience. As in a physical landscape, whether a certain feature deserves its own name depends on us and our relationship to the feature: is it useful to be able to refer to it with a name? Such namings are thus instrumental, but nevertheless refer to some real feature of the world. An 'ideal' language, one that captured every feature of the domain, would be unwieldy, since it would require a word for each point in the space, like using a 1:1 ratio map. Lewis Carroll in 'Sylvie and Bruno Concluded' (which inspired George Luis Borges' story 'On Exactitude in Science') saw the ridiculousness of the idea, imagining a kind of competitive map 'race' to produce the most accurate map, ending with a map of 1 mile to 1 mile, which became useless, resulting in people using 'the country itself, as its own map, and I assure you it does nearly as well' (Carroll, 1895). (There's an interesting parallel here to later work in situated robotics, see §4.2 Representations.) Of course, we need a map that has enough detail for us to get around and find what we want, without getting in the way. In terms of the points of interest we include on our map, if those were microphysical 'essences' there would be no room for the ambiguities that come with descriptions in a language. For example, in the case of 'water,' isotopes of water are included, not because of micro-physical similarity (we could keep changing this very gradually and asking each time 'Is this still water?') but because of the descriptions we attach to what we call water (clear liquid at room temperature that can be drunk, etc.). On the other hand, it is questionable whether we would accept as water a substance that is H_2O but for some reason is pink and fluffy at room temperature (example from LaPorte (2004)).

At this point, as with all theories that attempt to give an account of how we engage mentally with the world, we need to ask how the theory applies to itself, since our minds and the kind terms we employ are themselves part of the world. It could be a criticism that maps are abstract representations separate from the world represented, and so not suitable to account for our embodied, immediate engagement with the world. My response is to observe how a skilful orienteer uses a map: it is part of an embodied, reciprocal practice, interpreting what is seen in the world and on the map in terms of each other, constantly reassessing on the move. Moreover, the topographical claim is not merely one about our conceptual structure, but also about the structure of the world; it is a claim about how our conceptual structures can be said to be accurate in a meaningful way more than pure pragmatism, by which I mean a pragmatism that does not account for *why* a belief is useful.

The topographical model is an attempt to include the truths contained in both of the main historical camps in the field of natural kinds (the realists, who say that our terms make the same distinctions that nature itself has made, and nominalists, who say that our terms are a function of the way we

perceive a reality that in itself has no real distinctions). This is supported by recent discussions about biological species concepts. The consensus had swung from realism (Aristotle) to nominalism (Locke, Frege), and back to realism (Kripke, Putnam). Kripke and Putnam revert to essentialism, 'denying traditional accounts that make the reference of terms a function of something like their Fregean sense… [saying there are] properties which nothing can lack and still be of the kind' (Mellor, 1977, p. 299). Mellor (1977) criticised the latter realists saying the essential properties alluded to in these theories are more a feature of our theories than the world itself. Putnam's thought experiment about water and 'twin water' aims to establish the anti-Fregean conclusion that 'water' can have different extensions for users who have in mind the same Fregean sense. But this assumes that what matters is microstructure (H_2O or XYZ). Couldn't we say rather that what we would actually have discovered in Putnam's scenario is that not all water has the same microstructure (as indeed we have with the discovery of isotopes)? As Mellor (1977, p. 304) says, there may be a 'division of labour' in deciding reference (we defer to experts), but 'it might be a Fregean labour for all that.'

Rather than kind terms picking out a discrete set of entities definable with necessary and sufficient conditions, the topographical model claims that what they actually do is delimit an (in some cases vague) area of an 'intensional topography' where the axes along which the topography is plotted are the various measures of similarity we find between things in the world. Despite the fact that the map we form with kind terms depends on our interests, perceptual abilities, and history, it is still a map of some territory. The validity of the map must be justified without some naïve notion of direct correspondence though, but rather in a pragmatic way akin to a process of selection. Not just the selection that has been part of our evolution, which must have led us to be able to pick out real things (to run away from or towards) or we wouldn't have survived, but also the selective processes at work in the progress of empirical science (cf. Quine). It is realist in the sense that if there is something right about the way we categorise things, then we can use a categorisation to 'project,' and this fact is explained by our having successfully captured something about the actual structure of the world. It is also realist like causal theory in that subjectivity is part of reality, and there is a real way in which history plays a part in how the world presents itself to us: we 'resonate' with the world.

There is an important difference in emphasis between this account and Quine's (1969) view of natural kinds as sets. Sets are extensional, thus, so are kinds. But on the topographical view, kinds are not sets, and the conditions for their application are intensional, in that we name those features that matter to us, and draw the boundaries around features in a way that makes sense to us. Having

a boundary makes it look like a set, but that boundary depends on us, and so is not purely extensional, even though it does, more or less, succeed in picking out a class of objects which share certain properties which we can use projectively in predictions and explanations. It is true that if we ascend semantically to the formal level, where we can discuss how words are used, rather than whether they are used successfully, then this view shares with Quine's the fact that a term's significance depends on its extension, but unlike Quine's view, mine is not behaviourist to the extent of saying this exhausts its significance.

This is consistent with the causal individuation of kinds: when we understand a kind, we understand its causal properties (even though we may have to arrive here by a process of induction). But a causal account has trouble dealing with similarity, which the topographical account can handle, since we can retain the subjective sense of similarity without rendering our account subjectivist in a way that would cut us off from real structure in the world. In a topographical theory of kindhood, the similarities between things in the world are seen as defining a kind of landscape, and our kind terms pick out noticeable features on a contour map of this. This captures the way that kind terms depend both on the contingencies of our interests and on the actual world. It describes 'joints' in nature, but exactly where we 'carve' them depends partly on us.

This is not anti-realist, since, given our epistemic limits and purposes, what is picked out by the terms is an objective fact. Where we draw the boundary is up (or down) to us; that doesn't mean the place where the line is drawn is not a real place. Once the criteria we are using are decided, the fact that something satisfies this is objective. Maps are made according to our conventions, but are constrained by the reality we are trying to portray. As long as we have the right axes, the completed picture will be a picture of reality, a better or worse one depending on the axes and our measurement of them.

For realists, a term refers to something only if that thing has certain essential properties, irrespective of the speaker's understanding of this aspect of the term's meaning or those physical properties. For nominalists, reference depends on a match between how something impinges on the senses and an internal model of the sensory impact of various kinds of things. Once the models are thus fixed, there is nothing, of course, to stop determinate parts of the world being objectively picked out. The topographical approach improves on causal realism by not discounting the role of the agent's understanding in fixing reference, and on nominalism by giving the way the world is a constitutive role in forming the categories we deploy through a process of diachronic feedback. Moreover, this feedback has a top-down effect on determining our sensory perception, and therefore the subsequent maps we draw (see §5.2 Feedback and Feedforward). We don't simply discriminate

between given experiences; we actively use our understanding to negotiate with the world, experimenting with the practical usefulness of different ways of seeing the world. Perceptions don't just lead to understandings, ways of understanding also lead to ways of perceiving.

The topographical model bridges the gap between inner and outer by building a model over time through engagements with the world. Rather than being an abstract re-presentation, the map is an active part of the process which is refined through use, leading to a situation where the map can, on occasion, replace the world as the former of perceptions and actions. The topographical model is constructed in a conceptual space, where there may be multiple dimensions depending on the various ways we find it useful to categorise things (e.g. size, colour, behaviour, chemistry, taste). Differences between concepts may not just be dichotomous or a matter of degree; there may be multiple discontinuities, some nested within others. Concepts should be analysed so as to be consistent with linguistic usage and useful for prediction and generalisation. The latter empirical considerations can trump the former ones, this being how language develops over time (Sloman, et al., 2003, p. 18).

1.5.1 Historical precedents

> A natural kind is like what in topology is called a neighbourhood…. Cats, for example, are like a star cluster: they are not all in one intensional place, but most of them are crowded together close to an intensional centre. Assuming evolution, there must have been outlying members so aberrant that we should hardly know whether to regard them as part of the cluster or not. (Russell, 1948, p. 461)

Russell did not develop these ideas, as he thought that there was no rigorous way to specify this 'closeness,' because similarity is too subjective a notion. However, the account I am advocating can cope with similarity without becoming problematically subjective, because the way in which things are similar, and the degree of difference that makes a difference, is negotiated over time. The structure is neither inherent in the world, nor a free-floating conceptual balloon, but rather is 'tethered' to strategic points.

For Millikan, members of kinds are similar to each other because they share a history, rather than a physical essence, and this shared history *causes* the members of the kind to be similar, thus grounding induction. The kinds of historical links that can produce these similarities are evolutionary: there must be a mechanism for copying traits in response to a certain environmental contingency (Millikan, 1999, p. 55). But purely historicist accounts like this fail in the important case of functional kinds. We can imagine two different organs in evolutionarily diverse lineages that function to pump a nutrient-carrying fluid around the body of their host organisms. To rule that we cannot classify both kinds of organs as hearts because they don't share a selective past seems arbitrary; the similarity of their function will allow projection and justify calling them members of the

same kind for many purposes. A possible response could be that we can take the organs as sharing a history without sharing an origin, in that both were selected for the same reason on different evolutionary lineages. However, that identifying of a shared selective circumstance is one that is only definable in terms of the function of the organ.

Of course, our explanatory interests might be such that there would be good reason to distinguish between different kinds of hearts, in terms of how they work, what they evolved from, and so on. Therefore, there may be no fact of the matter as to which is *the* best classificatory schema independent of our perspective, but if the possibilities do reflect the structure of the world in being constrained by it, then this is still a variety of called realism rather than conventionalism. Dupré's (1993) 'promiscuous realism,' which takes the entities of some domain and maps them onto a multi-dimensional quality space, resulting in clusters corresponding to groups of similar entities, is close to the account being advocated here: 'Dupré's account is realist because the clusters in quality space reflect the real structure of nature. It is promiscuous because there will be many different clusters on which we could choose to focus' (Cooper, 2004).

The topographical account, though, is less promiscuous, more faithful to nature. Not all classificatory schemas are equally deserving, and it could be legitimate to overrule a particular way of classifying, if one scheme is more apt for induction, or is simpler. Promiscuous realism, in other words, does not have the resources to justify normative judgements about conflicting classificatory schemas, and common-sense categories have equal standing to scientific ones (Dupré, 1996). If this were the case, then there would be no way to convince someone who was brought up to call a whale a fish that they should change the way they use those words. In effect we would have to accept that we speak different languages. But that is not only an unproductive approach to disagreements, it also doesn't fit with the historical fact that we have been persuaded to change the way we use words on the basis of evidence and the virtues of non-commonsensical ways of talking.

Another position in sympathy with the present one is the 'cluster kinds' of Boyd (1989). This is where higher-level kinds are a cluster of lower-level properties, none of which are necessary or sufficient, but which are clustered together due to some historical, causal reason, thereby supporting inductions. For example, biological species are part of a lineage that is inseparable from its spatio-temporal niche, the homeostasis being underwritten by the process of the exchange of genetic materials (Boyd, 1991, p. 142). This is a kind of functional rather than physical essence.

However, this is not a concept of kinds that has general applicability, being only suitable for non-basic kinds. Some kinds do have necessary and/or sufficient properties, for example certain, basic

physical kinds may necessarily have charge. Before seeing how a topographical account works when applied specifically to mental kinds, it is worth exploring how it handles kinds in another, far removed domain, since the account should have general applicability if it has any.

1.5.3 Pluto: A case study in natural kind term demarcation

> 'I'm fed up with all this 'bananas are not a fruit, tomatoes are' rubbish. If it can go in a fruit salad it's a fruit and scientists can go stuff it.' – David Mitchell

We may give microphysical essences the pride of place for natural kinds, but whether we do so is our decision, and we are only likely to do so when it comes to physical kinds. To say further that no group of things that do not share a microphysical essence can be members of a kind seems question-begging. Are there kinds of things that share causal natures without sharing microphysical essences? We may give causal definitions to the kinds we use in science, without specifying the physical causal processes that underpin them, but these will be, to some degree, vague and interest and/or context relative.

In 2005, The International Astronomical Union gathered to decide on the definition of a planet, after there had been debate about whether some newly discovered objects in the solar system were to be called 'planets' or not. After agreeing on the criteria of having a large enough mass to pull the object into a near spheroid shape and clear its orbit, they then say that 'nature decides' whether or not something is a planet: 'Our goal was to find a scientific basis for a new definition of 'planet,' and we chose gravity as the determining factor. Nature decides whether or not an object is a planet' (Prof Richard Binzel, The Independent, 2006).

On the face of it, it seems problematic to claim that nature decides when they made a ruling by majority vote. But, with a little more careful wording, the claim is not obviously wrong: we decide the criteria, and then it is a matter of objective, physical fact which objects satisfy those. But that does not mean that there are some things that are essentially planets in and of themselves. Not only is 'near spheroid' vague, and therefore what counts as near enough a matter of convention, but there are many respectable scientists who make reasonable arguments for Pluto remaining a 'classical' planet, and say that it is culture, not nature that decides such things.

> Language routinely recognizes natural categories that have no good scientific basis…. There's no way to define the lily that doesn't include a lot of tulips as well. And other words like shrub and weed don't have any kind of scientific definition at all. So why can't we just keep using planet however we… please? (Nunberg, 2006)

Such appeals to ordinary usage may be appealing, but are ultimately unconvincing in these cases, as ordinary usage should take into account the findings of science. Linguistic usage develops as part of the cultural evolution of our society, and should reflect our improving empirical picture of the world.

Actually, a new category was constructed to include Pluto: dwarf planet, or pluton. In terms of the topography of the conceptual space occupied by cosmological object kinds, it is a newly labelled 'hillock' on the side of the mountain that is the kind 'planet.' Since one of the 'axes of similarity' along which objects are placed is sphericalness, there will be some objects that are not objectively one or the other, but that should not lead to any kind of anti-realism about the categories; there will be many objects that do fall squarely into one or the other category, and as such we will be able to make valid inferences based on this categorisation. A criterion for a subset of a kind not being a kind itself, is if everything that is true of the subset is true of the broader kind except the single property that makes the subset different. For example, white dogs are not natural kinds because all the inductively useful properties that are true of white dogs are also true of dogs, except being white (example from (Machery, 2003)). In the case of dwarf planets, I take it to be different along a number of such dimensions and so to be separate kind. According to the topographical view, this is accounted for by the 'prominence' of the peak, and is open to debate, like the discussion among climbers and cartographers about what counts as a separate summit or a subsidiary peak.

Whether Pluto is a planet is, in my opinion, a question on which the jury is out; we will know when future users of the term have learnt and accepted a usage from the textbooks they are exposed to.[2] But the point is clear: although we may decide on a causal definition for kind terms, it is our decision where we draw boundaries between kinds, and the causal features we focus on will be relative to our interests. However, this should not lead us into nominalism, as the map is an ongoing project, dependent also on the real structure of the world, as exposed to us through our ongoing attempts to navigate in our environment.

1.8 Cognitive Kinds

What does this view mean for how we view the kinds of states referred to when we do science about the mental causes of behaviour in beings like us? In this section, I will make some initial remarks that will be expanded on (§4.1 Mental Kinds) after clarifying other important metaphysical issues.

Concepts are formed from non-conceptual parts through time, and it is only when we have managed to bring together these parts into a generalisable whole that we can be said to 'possess' a concept (see §4.3 Content). These non-conceptual parts can be images, sensations, bits of language, etc.,

[2] Since the time of writing, my infant daughter was given a rubber place mat that has the planets on it, and there were eight. So maybe she will grow up finding it natural to think that way. Pluto *was* there on the edge, but depicted as left out, and sad.

and the results of these accumulations can be grouped together as kinds using a topographical model of similarity.

We can give pretty clear conditions for the holding of explicit, linguistically stated beliefs. The case is not so clear when we are attributing beliefs in less explicit cases, like that of animals. In some sense the dog that learns how to open a door does have beliefs about that door handle, but not in the same sense most adult humans do: the dog probably learned an association between a behaviour and a desired result, rather than having an explicit belief about 'doorknobs' which can be used in a more inferentially flexible way. Thus, an explanation of the dog's opening the door by referring to its beliefs is in some sense instrumental; it doesn't have a concept of DOORKNOB that it can manipulate in off-line thought processes in the way humans do. But there are commonalities between a dog's belief and ours: they both fulfil certain conditions, like being caused by sensory experience of specific aspects of the environment, being used to drive action, etc. In terms of the topographical account, we could say that a dog's beliefs and ours are both part of the same 'hillock,' in that they share some characteristics but not others. The human concept is 'higher' in that it is apt to be used more flexibly in more types of cognitive process: that is, it supports more generalisations.

This fairly heterogeneous view of concepts, which places such a variety of mental particulars as dogs' and humans' representations of doorknobs under the same kind, is vulnerable to the objection that these particulars are too heterogeneous to count as a kind, since they only share the fact that they represent the same things, whereas kinds should share many more inductively useful properties. For example, Machery (2003) argues that there are at least three different kinds of concepts (exemplars, prototypes and theories) that are all concepts in that they are 'poised to be used in our higher cognitive processes' (Machery, 2003, p. 451), but which are used for different ends and 'possess different representational properties [and] different functional properties' (Machery, 2003, p. 457).

However, Machery, in my view, is focusing on a rather narrow set of inductively useful properties, that is, only those of interest to psychologists (Machery, 2003, p. 436). If we grant that different cognitive processes are involved in making inductions, categorising, etc., it still may be that there are properties of interest to others that are shared by these various sub-types of concepts. In effect, Machery does not go 'high enough' when he is talking about how concepts are used 'by default' in higher cognitive processes. Inferring a list of probable attributes objects that fall under a kind possess may call upon a different cognitive process than categorising a given object or making predictions about its behaviour, but when it comes to making a conscious decision to act that may, if you are trying to take into consideration as much information as you have at your disposal in order

to come to the best possible decision you can, and if the concepts used in such processes can be called a kind because we can form useful generalisations about them in terms of the kinds of behaviour they produce when combined with other concepts, then we have a case of a higher-level natural kind. Moreover, these higher-level kinds can exert top-down influence on our lower level concepts, a fact that Machery does not consider.

Machery explicitly rules out a view of natural kindhood being a matter of featuring in laws, even if those laws are taken to be context sensitive, *ceteris paribus* generalisations, because such 'laws' are missing the fact that 'at least one causal mechanism… accounts for these generalisations' (Machery, 2003, p. 447). Again, this is only to see such causal mechanisms as working in one direction. If the higher-level kinds, which are heterogeneous in terms of their lower-level causal origins, can nevertheless be part of the causal mechanism that explains intentional action, then there is a strong case for counting such cognitive kinds as natural kinds. It is a large part of the aim of this thesis to make this case.

Another way we can see that there is a top-down influence on lower-level cognitive processes is in the case of categorisation. If it is the case that there can be such top-down influence, it must be that the higher-level and lower-level concepts share more than just picking out the same objects in the world, and that there is a much more intricate and intertwined landscape of concepts, rather than the simply bottom-up, modular view of Machery. 'Whales are fish' was not wrong because 'whale' and 'fish' have always referred to what they do while the concepts WHALE and FISH referred to subjective categories. Categorizations are not just given by nature, but depend on our interests too. We have decided that 'fish' means more than 'a creature that swims with fins in the sea,' not because it is 'more natural,' but because it is more useful, satisfying, neater, fits our purposes better, and so on. Past speakers were saying something different; in a way, they were using a different language. This leads to a question: when meanings change, what is left of the old meaning? On a topographical view, the meaning has been added to, differently subdivided, additional features added or some taken away, more details resolved, focussed on or given more prominence, but it's still a similar enough 'lump' in intensional space for us to be able to communicate. And, although the way the world is 'carved up' may change, the categories can still be regarded as natural. As our language changes, due to cultural progress, we refine the concepts we use to categorise and make inferences depending on how useful they are, which in turn depends on the way the world is.

In the cognitive sciences, the obvious candidates for projectable terms that enter generalizations on the basis that they share causal properties are the folk psychological categories of beliefs and desires that enter intentional explanations like, 'A person who desires X and believes that doing Y is a way of

getting X will do Y (all other things being equal).' The question then is: is the autonomy of these explanatory schemas threatened by the fact that in a physical universe where there are no 'mysterious' forces at work, all 'special scientific' kinds must be realised by, and supervene on, physical kinds? In the next chapter, we will address this reductionist worry through a subtler understanding of what it is to be physical, to be reducible, and to be causally closed. This refined understanding will then be used to show how the account I am defending allows for an autonomous realm of psychological causation within a physical world.

Chapter 2: Physicalist Reductionism

The status of cognitive states like beliefs as natural kinds that can enter into causal explanations is threatened by reductive arguments purporting to show how such higher-level states are nothing more than a convenient shorthand for the complex causal explanations involving the fundamental kinds of the physical sciences. Before countering this threat head on, we should first make sure we are clear about the terms of the debate, namely, what does reduction involve, and what is the nature of this physical stuff to which everything is argued to reduce to?

2.1 Reducing Reduction

The status of psychological terms like beliefs and desires as kind terms in explanatory scientific statements is dependent on the ontological status of the intentional mental states they refer to. This status may be threatened by reductionist arguments that claim, in brief, that in a physical world, it is the causal powers of the matter that constitutes such states that do the actual causing. This debate can be seen as parallel to the individualism/holism debate in the social sciences, where it may seem to make sense to say that causal generalisations about social entities are entirely reliant on causal generalisations about the individuals who comprise those entities. In response, holists point to how the social groupings we belong to, and the positions we hold within them, influence us as individuals, which is a sort of 'downwards' causation, from social to individual. We can find significant regularities not only in the behaviour of individuals in certain kinds of positions in certain types of groupings, but also in the circumstances under which various types of groups emerge. Such social entities, then, have a causal existence; they are visible to the scientific method. But there is an asymmetry between the social and individual levels of description: the latter is necessary for the former but not the reverse. Individuals can exist without social grouping, but not *vice versa* (see §3.3 Supervenience & Realisation). This asymmetry is less strong than it may first appear, though, as there is something necessarily social about humans. From the point of view of cognitive science, we need to understand (at least) three complexly interwoven mutually influencing realms: biology and its interaction with the physical environment (this is the same for all organisms), and both of these realms' relationships with the social environment (this is true for social organisms). A full understanding requires all three aspects.

In the relationship between the physical and psychological there is also an asymmetry: the latter needs the causal resources of the former, and many would conclude (like the individualists) that this means that the latter entities are mere fictions of convenience, shorthand for the real explanations that are too complex to comprehend. Fictionalism about psychological states is advertised as a response to eliminativism, by making it meaningful to talk about them without making risky

ontological commitments (Daly, 2008). Sometimes fictionalism may be the correct stance, for example, with regards to numbers it makes sense to avoid committing ourselves to their independent existence and restrict ourselves to using them to make useful truth claims. The same may be true of laws of nature: the law of gravity may be usefully imagined to be a force that is pushing things around, and use this, in combination with air resistance, to calculate the trajectory of a falling object. I will argue for a more robust ontological commitment to psychological states, by showing both how they are useful, and how they come to be in the world with the causal properties they have. Like social entities, psychological ones have an informative history, a past that explains their present. Different pasts may lead to the same physical present, but unless you understand the past of this thing, and what it shares with the pasts of things we would count as similar enough to group under a kind term, then we will be able to make fewer meaningful predictions. Without such a robustly realist account of mental states, they will be too insubstantial to support other important aspects of being an autonomous agent with minds of our own (see §7.2 Mental Causation).

This is where I would distance my position from Dennett's (1987) 'intentional stance.' As Hutto (2013) argues, Dennett's 'mild-realism' is equivocation; if talk of mental states is only true in an instrumental sense, because the physical level of explanation is practically inaccessible, then it is a form of fictionalism, which leaves folk psychology as a kind of myth and pushes advocates of such a position down the slope to eliminativism (Hutto, 2013, p. 597). Folk psychological explanations are not necessarily confabulations: the fact that we can be wrong about our understanding of the reasons for our actions implies the possibility of sometimes being right: some narratives are factual (Hutto, 2013, p. 600). Where I would disagree with Hutto is that I hold that reason giving can be a true explanation of action even where the mental states being referred to are subpersonal states, as when we explain the actions of other people and animals (see §4.1 Mental Kinds).

The motivation for defending anti-reductionism within a physicalist ontology comes not just from an intuition that there is a level of 'intentional causation' distinct from aggregates of microphysical causal processes, but also from an aversion to the kinds of explanations of cognitive traits that refer only to physically 'respectable' facts, e.g. involving DNA or neurons. The fact that scientists claim to have found a gene that explains brain size does not eliminate the need for non-genetic explanations of the need for a large brain. That would be to make the mistake of only looking at synchronic causes (see §2.2.1 Causal Closure). The presence of the gene provides one kind of explanation for brain size, but then we have to ask for an explanation of why that gene spread so rapidly and widely through the population. The broader explanation would refer to the adaptive importance of tool use and large cooperative groups. The presence of tools and political groupings become

environmental resources and constraints that shape individual development and the evolution of the species (see §5.2.1 Evolution).

The idea of constraints is important, with the most important constraining factor being the nature of physical stuff itself. Matter constrains the possible course of events, but does not alone determine what, with those constraints, emerges. Constraints can partly cause new things to appear with properties that are projectable: e.g. flowing water will form whirlpools in certain conditions. These emergent phenomena are, of course, wholly composed of physical matter, which is describable in terms of physical causal laws, but which nevertheless exhibit causal properties not ascribable to the parts but only to the whole, e.g. that of being able to suck down floating objects. If the laws of physics are at base indeterminate, these properties may be genuinely novel, but, in the case of phenomena like that of fluid dynamics, they are still physical causal properties.

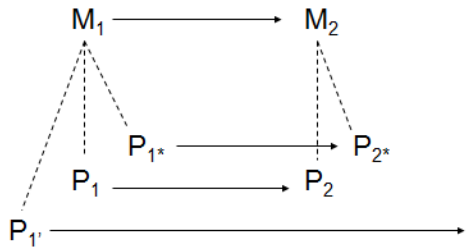
Reproduction and selection take such emergent properties and 'frees' them as kinds in their own right: they don't just emerge from the interaction of physical kinds each time like the first, but cause physical kinds to come together to form the next generation of those biological kinds. An indeterminate physical event, like that which happens when strands of DNA are recombined, causes the emergence of a macro-phenomenon, like an abnormally large brain. But then something more happens: that large brain causes the big-headed organism to be reproductively successful, and that strand of DNA becomes a gene that is inherited and reinforced by environmental constraints. This is life, and the causal properties of such living organisms are no longer describable using purely physical terminology. The same could be said of the mental kinds in psychological organisms. In these cases the new level of causation 'bubbles up' from the existing sea of forces and spreads outwards. As we will see in the case of cognition, the evolutionarily emergent objects of selection are complex, virtual machines (§6.1 Virtual Machines). Before we come to giving that account, though, we need to address the arguments given for and against reduction about mental states and statements that cite them in explanations of action.

The classic kind of reduction proposed by logical positivist philosophers like Nagel (1961) is analytical reduction, where 'bridge laws' act as translation principles between generalisations in the language to be reduced and those in the language of the reducing theory. This requires strong covariance, where (in the case where statements about X's are being reduced to statements about Y's) each Y is sufficient for X, and X is sufficient for the disjunction of Y's that may realise X. This account raises many questions, like whether disjunctions of properties can be kinds (Armstrong, 1978), whether disjunctive reductions should be barred, or whether general facts supervene on particular ones. Davidson's (1980) anti-reductionism instead targets nomological reduction. Here, rather than asking

if statements can be translated, nomological reduction requires that the empirical laws of one theory can be accounted for by the empirical laws of another (Davidson's argument being that logical positivism's failure shows the impossibility of providing analytic reduction, and secondly, since the mental realm is anomalous, that there are no laws to be replaced in a nomological reduction – see §2.2 Physicalism). We will return to these questions after a detailed presentation of the debate on the status of the so-called 'special sciences' between Fodor and Kim.

2.1.1 The 'Special' Debate

In the 70's, Fodor argued for the autonomy of the 'special' sciences (Fodor, 1974), these being sciences with laws that are not re-stateable in the laws of the sciences that govern the behaviour of the matter out of which the objects of these sciences are composed. The standard account of reduction at the time was Hempel's, where a law at one level can be said to be reduced to a lower level when 'bridge laws' are found that allow a 'translation' from the language of the higher level to that of the lower. So, for example, it could be claimed that statements in chemistry about the properties of the elements and their combination can be stated in the language of physics, mentioning atoms, electrons, etc. Fodor argued for the autonomy, i.e. non-reducibility, of the 'special sciences,' meaning non-physical sciences like psychology or sociology, on the grounds that such sciences contain natural kinds that are multiply and heterogeneously realised. He took a natural kind to be a class of objects which fall under a causal law as either antecedent or consequent. So, if we have noticed a causal regularity among a class of objects, and those objects do not share sufficiently similar physical characteristics for us to reclassify them as physical rather than special-scientific kinds (i.e. they are multiply realized), then there is not a physical level law to which the special-scientific law reduces. It can be illustrated with this diagram:



The arrows represent causation and the dashed lines realization. It illustrates cases where there is a causal relationship between two non-physical kinds, and where these may be physically instantiated in a variety of ways. Further, although there may be causal laws connecting the instantiating bases, these laws are between kinds that do not fall under a single class at the physical level, or there may

be no physical law at all. This heterogeneity at the level of the realization bases means that bridge laws between higher and lower-level phenomena cannot be found, since bridge laws should be biconditional (i.e. given a bridge law it should be possible to infer M from P and *vice versa*) whereas causal laws are simple conditional statements (i.e. given M_1, M_2 can be inferred). Thus, reduction via bridge laws is blocked due to the inability to infer the realization base given the realised higher-level kind. This is because the higher-level kinds are realised in ways that are insufficiently similar. It could be argued that there are species-specific bridge laws, where the realisation bases would not be 'wildly' heterogeneous. However, it seems to me that even within species we wouldn't want to tie ourselves down to particularly narrow definitions of realisation bases. If the various kinds of physical states that can realize the non-physical kinds are 'wildly' disjunctive, and as such do not form a kind, they would not share many properties apart from the fact that in certain circumstances they happen to realize certain non-physical kinds.

Fodor illustrated this disunified picture of the sciences with examples such as oxbow lake formation in geology and money in economics. Oxbow lakes are formed by rivers despite differences in the mineral make-up of the land through which they flow. We can make meaningful generalizations about money despite the fact that money can take the form of a bank-note, an electronic trace, or a conch shell. In response it could be said that in the case of oxbow lakes, the real dynamics are happening at the physical level, that our statements about oxbow lakes are mere useful shorthand for a very complicated story, but that if that story were told, rather than a reduction we would have an elimination, since we could predict the same events (better, given no restriction on computational power) despite the heterogeneous realization of what we call oxbow lakes. In the case of money it could be said that what plays the causal role is our beliefs about money, not the stuff we accept as money due to the conventions of the society we live in. Here we have meandered into the area of beliefs and their consequences, of explanations that cite such phenomena as beliefs as antecedents in causal generalizations. Are they like oxbow lakes: rough and ready shorthand for the impossibly complicated true causal story of neuronal firings, chemistry and ultimately physics; are they things that only exist in the stories we tell each other and ourselves about our behaviour, or is there a robust sense in which beliefs, desires and other intentional mental states play an irreducible causal role? That is, are intentional mental states natural kinds in that they have causal natures of their own?

Fodor says reductionism implies 'every natural kind is, or is coextensive with, a physical natural kind,' where natural kind predicates are 'bound variables in a science's laws' (Fodor, 1974, p. 690). However, the 'special sciences' make interesting (i.e. counterfactually supporting) generalisations

'about events whose physical descriptions have nothing in common,' e.g. money and psychological states (Fodor, 1974, p. 691). Thus, if higher-level kinds enter into laws but reduce to disjunctions of physical kinds, then the reducing 'laws' will have disjunctive antecedents and consequents, which may be logically equivalent to the original law, but will not themselves be laws, since their bound variables will not be natural kinds, as they will not be projectable (Fodor, 1974, p. 695; Kim, 1992, p. 520). The upshot is that we must either give up the idea of higher-level kinds, or give up reductionism and accept a token physicalistic ontology implying an event/property dualism (in contrast with substance dualism) (Fodor, 1974, p. 689). Fodor urges us to towards the latter.

Kim, on the other hand, argues that, given the principle of the causal individuation of kinds, reductionism is unavoidable: the realisation bases could not be 'wildly' heterogeneous, since then the higher-level kinds would not be sufficiently homogenous to form a kind (Kim, 1992). He argues that if special-scientific kinds supervene on and are realised by physical ones, then, given the causal closure of the physical domain, the causal power of the 'higher' kinds comes from that of the lower. Thus, there is reduction to each realisation base, so no higher-level kind in fact (Kim, 1992, p. 523). To deny this would be to accept some sort of downward causation (violating causal closure) and 'magically' emergent powers. Since the physical is causally closed, and events cannot be 'overdetermined,' all causal properties are inherited from the realiser, and non-physical kinds cannot have a causal nature to call their own.

Here are the main steps in the argument, which will be expanded on below (Kim, 1992, p. 136):

1. Events are instances of property instantiation.
2. Emergence: for a causal property to belong to an autonomous realm of investigation, its causal powers must be novel.
3. Supervenience: all properties are physically realised and dependent on physical properties.
4. Downward causation: given emergence and physicalist realisation, for one event to cause another, it must also cause its realisation base.
5. Physical causal closure: every physical effect has a sufficient physical cause, so the realisation base of the caused event has a sufficient physical cause.
6. Exclusion: given the presence of a sufficient physical cause, there is no room for another, 'downward' cause.

Kim uses the example of jade as a purported natural kind, multiply realised by two distinct chemical kinds, his conclusion being that jade is not a natural kind because the similarities between the two

kinds are accidental, in that they are separately explicable by reference to the respective chemical kinds.

In response, Fodor accuses Kim of picking a fight with a straw man. Indeed, jade is not one kind but two, and so not projectable. But functional properties like pain *are* projectable, and are multiply realised. He makes a distinction between multiply-based properties that are *disjunctive* and multiply-based properties that are disjunctively *realised* (Fodor, 1997, p. 153). Kim 'wants to *just stipulate* that the only kinds there are are (what he calls) *local*.... In effect, Kim wants to make it true *by fiat* that the only projectable kinds are physically homogeneous ones' (Fodor, 1997, p. 161). We will return to the issue of locality later (§3.1 Causation).

The exclusion argument is analysed in Humphreys (1997b, pp. 1-2), looking at a similar argument that assumes mental events are not physical events and concludes that mental events must be causally irrelevant:

1. If an event X is causally sufficient for an event Y, then no event X* distinct from X is causally relevant to Y (exclusion). (This needs to be supplemented with a synchrony condition.)
2. For every physical event Y, some physical event X is causally sufficient for Y (physical determinism).
3. For every physical event X and mental event X*, X is distinct from X* (dualism).
4. So: for every physical event Y, no mental event X* is causally relevant to Y (epiphenomenalism).

In these arguments, some technical vocabulary is carefully defined, but some equally important concepts that are from everyday language are left somewhat vague and unanalysed. A key word is 'event,' which we may think is unproblematic since we all know what events are. However, a lot of work is done by this word, and our everyday understanding needs careful examination, which is the task of the next section.

2.1.2 Stuff Happens

An event is something that happens. Classically, they happen to objects (apples fall, fallings don't apple). But some would say objects are events, in that they are not static and unchanging, but are happenings at a time and space. Or are events just facts? And, how 'fine grained' are events?

Kim relies on a 'property instantiation' account of events. Events are the coming to be of causal properties at a particular place and time: 'We may consider talk of 'event kinds' as equivalent to talk of 'properties' of events, since every property of events can be thought of as defining a kind of

event, namely, the kind comprising events with that property' (Kim, 1996, p. 59). When something happens, what happens is that a certain (causal) property is instanced. It is these properties that are then plugged into his argument. However, things might not be as simple as he makes out.

Equating events to instances of properties is to conflate facts and events; describing things differently results in different facts. Kim says 'Brutus stabbing Caesar' and 'Brutus killing Caesar' are different events (Kim, 1976). But did Brutus, after stabbing Caesar, have to perform another action, namely, killing him? After all, there is on the face of it only one action, and so one event, being described differently. Facts are too fine-grained. Kim tries to side-step such attacks by distinguishing between 'constitutive' properties and ones referred to in 'extrinsic' descriptions (Kim, 1976, p. 318). This may prevent there being ridiculously many different events happening at the same time and place, but the thrust of the objection remains: to count this instance of stabbing as one event and this instance of killing as another is to count too many events. There may be different explanations for why Brutus stabbed Caesar and for why he killed him, but explanations are geared to propositions rather than to what the propositions are about.

Compare Davidson: 'It is not events that are necessary or sufficient as causes, but events as described in one way or another…. Events are identical if and only if they have exactly the same causes and effects' (Davidson, 1969, pp. 301-6). So, the causes of stabbing are different from the causes of killing in that a knife is necessary for the former but not the latter. That is to bring in a counterfactual element, though: the knife is not necessary because Brutus could have used poison. However, here we are talking about a factual event in the actual world, which can be variously described as stabbing or killing. Different descriptions do entail different explanations, because the description directs our interest, and so affects what is relevant, in a particular way. We can talk about the different causes of stabbing and killing, so either we accept that they are different facts, or we say that various possible causal descriptions attach to an event and give it its identity. Such complex entities would not be very useful as a basis for generalisations though, as each event would be unique. Either way, it is the causal properties named in descriptions that can form the basis of the general statements we are interested in.

Kim could also be attacked for his reliance on a property instantiation account of eventhood. His reductive conclusion doesn't follow if we take events to be 'bare' occurrences, which may be described in various causal ways depending on our interests. For example, the shot that killed Archduke Ferdinand could be described as the event that started the First World War, or as the event that ruptured his heart, and so on. Of course, given that we want to be realists about the world we are describing, the structure of the occurrence constrains the possible descriptive

properties we can clothe it with, but that doesn't mean that events are property instantiations as there is a many to one relationship between properties and the events they describe, unless an event is identified with all the possible properties that could be used to individuate that event. But descriptions cannot have such ontological clout; rather they are like indexicals: they point to an event, rather than constituting it.

But can't we ask, 'What is it about that event that makes it the event that we take it to be?' That is, what gives that event the identity it has in our world? There are many ways it could be described, but not all these are of the same status, since some of them are necessary in that they would be found in all conceivable, nearby possible worlds (nearby in that they resemble the actual world enough for us to be able to say that it was the same event). Physicalism is committed to there being only one kind of stuff, not one kind of thing. What counts as the same depends on the description it is under, of course. In terms of being the event that caused the First World War, it wouldn't matter if a different kind of gun were used. In terms of being the firing of a 1910 Browning, changing the gun would make it a different event. Thus, every change could make an event a different event under some description, so events are not independent of the properties used in our descriptions of them. What counts as an event of a particular kind depends on the kind of explanation we are going to use the event in. It makes no difference to the explanation of the causes of the First World War that it was a 1910 Browning. It does make a difference that it was a Serb nationalist who pulled the trigger.

So, we can separate those properties that are essential to an event under a particular description by seeing what can and cannot be varied while retaining its explanatory power in the intended explanation. If in all the possible descriptions of this event that could be called the same event the intentional facts are present, then these are necessary to the identity of the event. And if these are necessary (and reductionism is false) then the physical parts of this description cannot be sufficient. Indeed, they might not even be necessary (except in that there must be *some* suitable physical facts: he couldn't have killed him with a peach). If the Archduke had been stabbed by a Serbian Nationalist, it would still have been the event that started WW1. If he had been shot by a psychotic Austrian, it would have been a different event. But if we are interested in explaining what killed the Archduke, the wielder doesn't matter, the weapon does. So, events can be seen as instantiations of certain 'core' causal properties, and Kim's argument might still go through on these, since if these properties include mental properties, there is still going to be a conflict between the causal powers of the event as described mentally, and the causal powers of the physical supervenience base of those mental properties.

Either events are unanalysed things (one thing happened when the Archduke was shot, but it can be described in different ways, as his assassination or as the start of the war), or events have structure (two things happened but they have the same physical-level description). Can we really think of unanalysed events though? How do we enumerate them? Can we point and say, 'That is an event' without giving some description to specify what aspects of the world are being pointed at? Without the descriptions by which we categorise, there is just an undifferentiated and unbounded boiling sea of change.

Putting the potential vulnerability of the property instantiation account of events aside for the moment, Kim's argument also relies on a causal theory of reference of kind terms, where kinds are baptised by acts of 'original ostension' (Kim, 2005, p. 113), the term referring to all and only those objects that share essential microphysical properties with the original object (Kripke, 1972). But, as pointed out above (§§1.2 Are Natural Kinds Found or Made?, 1.3 Rigidity), baptism doesn't work by ostension alone (LaPorte, 2004); some description is necessary, and the extension of kind terms is also decided partly by convention. However, perhaps the reductionist argument can be recast in terms of the causal properties used in explanations that refer to events as described. On that assumption, let's look at the other premises in Kim's argument.

For the exclusion argument to work, events need a strong sense of causal identity, which may be given by supervenience if we can identify a higher-level property instance with a collection of lower-level property instances, that is, where the causal properties of the supervenience base can simply be added up to give the causal properties of the supervenient state (Humphreys, 1997a). However, this assumes that each individual part of the whole is still a separate entity, and they are all acting together at each moment (if supervenience holds synchronically), rather than those individuals now being subsumed in a larger whole in such a way that they no longer can really be said to exist as individual objects. This stronger sense of 'fusion,' where individual parts no longer have separate identities as such, is a variety of emergence. Causal closure may be maintained diachronically (see §2.2.1 Causal Closure), as we could trace a causal chain back to a time when the parts were involved in independent events. If, however, we broaden supervenience to allow a temporally extended supervenience base, then it is doubtful that supervenience can still do the job Kim requires of it.

The key to Humphreys' argument is the notion of 'fusion,' by which he means a real physical operation by means of which separate physical entities become parts of 'a unified whole in the sense that its causal effects cannot be correctly represented in terms of the separate causal effects' (Humphreys, 1997b, p. 10) of the constituting entities. This is because the causal properties that some of those entities had as individuals have been 'used up' in the process of fusing. Thus, it is a

mistake to see microphysical parts making a causal contribution to a whole. In a way there is a bi-directional set of constraints: the properties of the whole are constrained by the nature of the composing parts, and, at the same time, the properties of the parts are constrained by the whole in which they partake.

We could talk about the causal properties of the individual entities that compose to make the larger entity, but those causal properties are not the same as those they had before; they are now determined in part by context, by being a part of that whole. So, to speak of each part having its own causal properties, which are added up to form the causal properties of the whole, is to miss the fact that the parts have those causal properties because of their context. This is always true in a sense; causation is never context free, and *ceteris paribus* clauses can't cover up that untidy fact. In the case of such fusion, the context that is the macro-entity takes its parts with it; it is an entity because it persists, and because it persists (and takes its causal properties with it) we can make predictions about it. Of course, we can choose to focus on the causal properties of the parts, and there may be good reasons for doing so sometimes (e.g. in investigating a disease), or we can choose to take it as a whole, or indeed as part of a wider entity (e.g. a population). But that doesn't mean that calling an organism an independent entity is arbitrary, as it is the locus of complexly interweaving causal processes, knotted ('fused') together. It is natural to take such an entity as independent because of the amount of explanatory work that this can do.

For Lowe (2008a), talking of events as having causal properties is misguided. Lowe advocates the view that all causation is substance causation, that is, the causing of events by substances: 'causation is fundamentally a matter of substances exercising their causal powers to act upon other substances possessing suitable causal liabilities' (Lowe, 2008a, p. 164). For example, a magnetized piece of iron acts upon some nearby iron filings to make them move towards it. Talk of event causation is not wrong, just not ontologically fundamental: 'Events consist in the doings of substances.... They are mere changes in things and not the source of those changes' (Lowe, 2008a, p. 342).

By substance, Lowe means an '*individual* substance': an ontologically independent entity that bears properties, stands in relations to other substances, persists through time and undergoes qualitative change over time. The causal powers and liabilities of these substances are a species of disposition, 'intimately connected with the *natures* of those substances, with what *kinds* of substances they are and how they are constituted' (Lowe, 2008b, p. 166). What he calls 'the philosophers' myth of event causation' is probably a hangover from positivism's insistence that 'science is only concerned with recording observable events and noticing their patterns of recurrence, [whereas] real science is

concerned… with revealing the causal mechanisms underlying and explaining recurrent patterns of events' (Lowe, 2008b, p. 166). When it comes to what constitutes us as substances known as human agents, he rejects the 'Humean' view of persons as 'bundles of perceptions.' We are not constituted by mental events, but are substances that 'have' those events. So, humans are 'psychological substances,' and as such have causal powers.

If we take causation to happen between 'property instance atoms' (Humphreys, 1997b, p. 13), which are properties of a psychological subject ('substance' in Lowe's terms) and do not decompose unless they are interacted with in such a way that they are no longer parts of the causal chain, then the parts of which they are composed are taken along for the ride as parts of that emergent individual. This rejects the assumption that the only way to cause a higher-level property is to cause its supervenience base. Macrophysical as well microphysical entities have irreducible causal properties, since, if, when looking for causes, you only see microphysical causes, then you will miss much of what is happening; I have a causal history going back to my birth, at least, even though none of the atoms in my body may be the same as then, but, if you think the only really existing things (as in causing things to happen in the world) are microphysical particles, then I do not really exist. This is a conclusion I would like to resist.

In response, a reductionist could say that while it might be useful to describe a rock as a whole in order to make predictions, given enough computational power it would in theory be possible to make the same predictions, or even better (whether or not determinism is true) without making reference to the 'rock.' In the case of an organism, however, the level of description that takes it as a single instantiation of a natural kind is, I would claim, the 'best' level of description, in that the lawlike statements that those kind terms enter into do not reduce to (their predictive power is not 'trumped' by) statements that only involve kind terms from physical science (whereas lawlike statements about rocks perhaps do). The feedback dynamic inherent in biological evolution, which gives evolved organisms their special place in the ontology of the universe, is also found in the realm of cognition, where there are the evolution-like processes of social evolution and individual development. In the case of animals like ourselves, which develop in an environment where language is a vital developmental resource, and where self-ascription of mental properties allows social and personal 'evolution' to feed off each other (due to, for example, expectancy effects in selective perception), the process of naming mental kinds itself (through scientific endeavour) affects the kinds that emerge (see 5.2 Feedback and Feedforward).

If there are such causal properties, then they will be visible to science. Higher-level kinds are realised by collections of lower-level ones, and so there is always a description of what happens at

the physical level. However, the explanation of the higher-level kind requires an evolutionary-type story, where the explanatory role will be played by the higher-level kind not the lower-level descriptions. These matters will be explored in detail below (Chapter 3: Levels of Causal Explanation).

2.2 Physicalism

For something to cause a physical event, it must itself be physical. This is in essence the principle of the causal closure of the physical used in Kim-style reduction arguments. It seems innocuous on the face of it, which is probably why Kim doesn't argue for it, assuming it is something that all good physicalists would accept (Kim, 1992). However, under the surface of this principle are some assumptions worth unearthing. Causal closure and causation will be the focuses of the next two sections; here I will look at the grounds of physicalism itself.

What is a 'good' physicalist? I take it to be someone who doesn't resort, or tolerate others resorting, to mysterious, spiritual, or otherwise non-physical forces or entities to explain the existence of the mental, whatever that is. That verges on being circular and trivial, though it will serve as a starting point. This is not the place for a survey of the dualist debate, but clearly there is something non-trivial about saying that the mind is physical. If someone claims that their intuition is that the mind is physical (because they have been brought up as 'good' physicalists), the fact that they are right doesn't mean that they know what they are saying (Democritus may have been right that all matter is composed of atoms, but he didn't know what that meant). It is a mystery how my experience of being in the world could be a purely physical thing, just as my cup of tea is, even though there is (probably) nothing it is like to be a cup of tea. All the atoms in my tea could be described, and there would be nothing more to know about what it is to be tea. But most (unindoctrinated) people feel that a description of them down to the atomic level would somehow miss something essential, namely, what it feels like to be them. Kim asks the question like this: 'How can there be such a thing as consciousness in a physical world, a world consisting ultimately of nothing but bits of matter distributed over space-time behaving in accordance with physical law?' (Kim, 2005).

The physicalist position (or rather, the physicalist methodological precept), often takes it for granted that we know what we mean by 'physical,' but closer interrogation may reveal some 'smuggled' assumptions. Berkeley defines it as 'an inert, senseless substance, in which extension, figure, and motion do actually subsist' (Berkeley, 1710, §9). He shares this assumption that it is senseless with

Descartes and his followers. But, why assume this? Panpsychists challenge it (see §6.3 Panpsychism & Composition), and even if we don't, it is instructive to ask why we assume it, and furthermore to state the conditions under which something composed wholly of matter could acquire sense. Physicalists who believe that there is something about the arrangement of the matter in bodies like ours that make them suitable to experiencing sensation also disagree with Berkeley, but not in the same way as panpsychists (see §3.4 Emergence).

Physicalists like me, then, are committed to there being only one kind of stuff in the world, not one kind of thing. That is, there is more to us than particles following physical laws, but that something more is not from outside the physical world, or in any sense separate from the physical. I assume that no physical laws are broken in building things with minds, and that a purely natural explanation can be given for the existence and nature of our mental properties. There may, then, be at least two senses of what we mean by 'physical property' when we say that all properties are either physical properties or supervene on physical properties. One is that they are the kinds of properties that physical theory tells us about, the other is that they are the kinds of properties that 'paradigmatic' physical objects are made up of. Moreover, although these may be the same set of properties, they also may not be (Stoljar, 2009).

Stoljar also subtly distinguishes the completeness question (What does it mean to say that *everything* is physical?), from the condition question (What does it mean to say that everything is *physical*?), and argues that we should focus on the latter: 'If Thales says that everything is water, or Up-to-Date-Thales says everything supervenes on water, we don't understand what he says unless he says something about what water is. The physicalist is in the same position' (Stoljar, 2009).

If we are tempted to define the physical (by define I mean doing more than giving an ostensive definition *a la* causal theory) by referring (deferring?) to physical theory (theorists), then we are stuck by Hempel's dilemma: we can either define the physical with regards to current physical theory, or to some future completed physical theory; the first horn is unattractive given that we can assume our current theory is in many ways incomplete and mistaken; in the second case it is trivial, since we don't know what the completed theory will be like, so we can just say it will include whatever properties we find. Kim openly opts to take the second fork: '…biconditional laws would allow the rewriting of the laws of the reduced as laws of the reducer, and if any of these rewrites is not derivable from the pre-existing laws of the reducer, it can be added as an additional law' (Kim, 1990, p. 18). However, given that neither option is acceptable (plus the logical fact that there are no options between them), we should conclude that we cannot define physicalism by reference to physics.

Recently there has been an interesting trend in metaphysics towards rejecting the kinds of physicalism that rely on causal arguments of the kind Kim utilises, e.g. Ney (2016). These are labelled neo-Russellian, as they are similar in spirit to Russell's position of neutral monism, which claims that although there is a single kind of substance, which can manifest either physical or mental properties, it is not itself characterisable as physical. These neo-Russellian positions reject arguments from the causal completeness of physics to the exclusion of mental causes since there are in fact no causes at the microphysical level. One of the main arguments against microphysical causation is that generally accepted microphysical laws are 'time-reversal invariant,' whereas causation requires there to be a temporal asymmetry: causes happen *before* their effects. It is only when microphysical parts are gathered into wholes that can be lumped together in multiply realised kinds due to their trajectory in the direction of increasing entropy that causal statements can meaningfully be made (Ney, 2016, p. 148).

I will not be investigating such positions further here. Suffice it to say that if a form of neo-Russellian physicalism turned out to be correct, it would help rather than hinder the case for non-reductive physicalism, as it agrees that mental causes are not excluded by physical ones, and that causes are correctly seen as applying to objects and states composed of physical stuff, the causal properties of these being constrained by, rather than determined by, the nature of the composing matter and its context. Furthermore, the idea that it is multiply-realisable objects in dynamic, diachronic processes that are the loci of causal statements resonates with the account being defended here.

Chalmers (2002) includes another kind of Russellian position in his taxonomy of types of physicalism. Panprotopsychism is the idea that in order to explain the presence of consciousness in a physical universe, the physical ultimates out of which everything is composed must have some properties apt to lead to consciousness, properties that are not captured by physical theory, for, as Russell pointed out (Russell, 1927), physics talks only of the relational properties between bits of matter, but is silent on what intrinsic properties these bits might have. We will return to these arguments below (§6.3 Panpsychism & Composition).

Another species of physicalist in Chalmer's taxonomy is characterised as accepting interactionism, emergentism and property dualism, and rejecting physical causal closure. Chalmers calls this a type of dualism, in that it is a property dualism, but it is still a kind of substance monism.[3] I will defend a version of functionalism that fits this category, Virtual Machine Functionalism, which has the benefit of being a valid intermediate position between type and token physicalism. Type physicalism claims

[3] According to Chalmers, such Type-D dualists include Foster 1991, Hodgson 1991, Popper and Eccles 1977, Sellars 1981, Stapp 1993, and Swinburne 1986.

that for every distinct higher-level type (like a kind of mental state) there must be a corresponding type of physical state. However, this has the consequence of our not being able to talk about humans sharing kinds of mental states with creatures who are physically dissimilar. Token physicalism on the other hand merely requires that each instance of a higher-level kind is identical with some physical state, without requiring any commonality between these realising states. This permits functional definitions of mental states to be given, although some would question whether such states are indeed respectably physical (e.g. Jackson (1998) calls functional properties 'onlooker properties' rather than counting them as physical). It is this kind of 'looseness' in the ties between the mental and the physical that opens the door for Kim-style reduction arguments, as there has to be some relation between the kinds of physical states suitable for realising mental states. There are other possibilities between type and token, though, loose enough to allow in the kind of functionalism being advocated here, but not so tight as to lead to reductionism (c.f. Cussins' construction constraint (1990, pp. 374-8)).

Another well-known argument for physicalism which utilises some not dissimilar assumptions to Kim's is Davidson's (1970, p. 116). The assumptions are:

1. The principle of causal interaction: 'at least some mental events interact causally with physical events.'
2. The principle of the nomological character of causality: 'events related as cause and effect fall under strict deterministic laws.'
3. The anomalism of the mental: 'there are no strict laws on the basis of which mental events can be predicted and explained.'

The argument goes like this:

a) Mental event ***m*** causes physical event ***p***. (by 1)
b) Under some description ***m*** and ***p*** instantiate a strict law. (by 2)
c) Strict laws must be physical laws, not psychophysical laws. (by 3)
d) If ***m*** falls under a physical law, it has a physical description. (by definition)
e) If ***m*** has a physical description, it is a physical event. (by definition)
f) ***m*** is a physical event.

Davidson's account is ultimately unsatisfactory, in my view, because of the claim that mental events are strictly anomalous: if there are to be meaningful scientific explanations about human behaviour

that refer to mental causes this cannot be the case. He assumes that the mental is fundamentally different from the physical, then shows that the mental is physical if it is causal (see §3.3 Supervenience & Realisation). The nature of causation, laws and their relation to the mental will be investigated further below (Chapter 3: Levels of Causal Explanation). For now, we will work with a version of the standard argument for physicalism from Lowe (2000):

There are 3 assumptions:

1. A physical causal closure principle (see next section).

2. At least some mental events are causes of physical events.

3. The physical effects of mental causes are not causally overdetermined.

Conclusion: at least some mental events are identical with physical events.

Now, let's look closely at the first assumption. A lot rides on this principle, but, as mentioned, it is often taken for granted.

2.2.1 Causal Closure

We've seen the principle of causal closure of the physical (henceforth PCCP) used in an argument for reduction (§2.1.1 The 'Special' Debate) and for physicalism in general (§2.2 Physicalism). It is often assumed that any right-minded physicalist should accept this principle without argument, but if it is used in an argument for physicalism, that is question-begging. We should ask what form this principle takes precisely, whether there are different possible understandings of it, and whether it is the same principle being used when arguing for reductionism and physicalism. I will claim that the principles adverted to in the above arguments are distinct, and that the one used in arguing for physicalism is acceptable, while the other one is not.

Lowe analyses the different interpretations of the PCCP, some being stronger than others, and concludes that if it is too strong then it is question-begging, but too weak and it fails to establish the conclusion. His argument is that there are 'various forms of naturalistic dualism, of an emergentist character, which are perfectly consistent with the strongest physical causal closure principles that can plausibly be advocated' (Lowe, 2000, pp. 572-3). If a PCCP is to be used as a premise in an argument for physicalism it shouldn't be so strong as to stipulate the desired conclusion, but if it is weakened (e.g. to a diachronic principle – see below), then it allows for the possibility of non-physical causation. As an example of a weak version Lowe gives:

1) Every physical event which has a cause has a sufficient physical cause (Lowe, 2000, p. 575).

This is not strong enough to rule out naturalistic emergence because of the transitivity of causation (the physical cause could be a historical one). The stronger version that Kim needs in order to get his conclusion should be synchronic:

2) At every time at which any physical event has a cause, it has a sufficient physical cause (Lowe, 2000, p. 576).

Causal closure principles that don't include a stipulation that the causes must be concurrent with the effects can allow for diachronic emergence and downwards causation, given the transitivity of causation (Lowe, 2000, pp. 575-576).

For completeness, it's worthwhile listing the variations on PCCP that Lowe distinguishes:

(1A) All physical effects have sufficient physical causes. (D. Papineau)

(1B) All physical effects have complete physical causes. (D. Papineau)

(1C) Every physical effect has a fully revealing, purely physical history. (S. Sturgeon)

(1D) Every physical effect has its chance fully determined by physical events alone. (P. Noordhof)

(1E) No physical effect has a non-physical cause. (S. Sturgeon)

(1F) Every physical event which has a cause has a sufficient physical cause.

(1G) At every time at which any physical event has a cause, it has a sufficient physical cause.

(1H) Every physical event contains only other physical events in its transitive causal closure.

Lowe doesn't go into detail about the first four, but the discussion of the last four can be applied to them.

He says (1E) is too strong, since it renders redundant the non-overdetermination clause in arguments against downward causation. (1F) on the other hand is too weak in that in conjunction with premises (2) and (3) (of the above argument for physicalism) it doesn't entail the conclusion because it doesn't take into account the transitivity of causation. (1G), which he says is close to Kim's formulation, also fails to rule out nonphysical causes, since it doesn't rule out the possibility of simultaneous causation, where a physical state P has a sufficient physical cause, P', which causes P in part by causing a mental state M, which is also an immediate cause of P in that 'it is not the case that in the absence of either one of them P would still have occurred' (Lowe, 2000, p. 577). This sounds like he's taking P' and M to be individually necessary for P, which is something I would endorse, but since P' is (synchronically) sufficient for M, and given the transitivity of causation (which is a

conceptual principle, not just a temporal one), then P' is sufficient for P, and M is epiphenomenal, in that it doesn't have an interesting causal role of its own.

Lowe, though, wants something stronger than (1F), proposing (1H), where by 'transitive causal closure' he means the immediate causes of P, the immediate causes of these causes, etc. He says it is weaker than (1G) 'in that it does not imply that any physical event has a sufficient physical cause' (Lowe, 2000, p. 582), and that this gives it an empirical advantage over (1G) in that it is consistent with probabilistic causation. Here, in order to give mental kinds a role, he makes a distinction between causal events and causal facts, where the latter are facts about what certain physical events cause: M causes it to be a fact that P' causes P, which doesn't violate (1H) because M isn't an immediate cause of P', although it is a cause of sorts.

I don't think this works though: if it is the case that P' wouldn't have caused P unless M also occurred, then M is a necessary condition for P, just as P' is. I don't think a clear distinction between causal facts and causal events can be maintained: how does M cause it to be the fact that P' causes P if it isn't a causal event with an effect that is also a causal event?

I take from the foregoing that a 'weak' PCCP, one which allows for emergence, is sufficient to make an argument for physicalism but not for reduction, and a strong PCCP, which can be used to argue for reductionism, is question-begging in terms of an argument for physicalism. In other words, there is no problem with a principle of the form, 'everything is caused by things that are made up of physical stuff and no other kind of stuff,' but there is a problem with principles of the form, 'everything is caused by things made up of physical stuff and whose causal properties are describable in terms of the causal properties of their constituent matter.' This can be seen in terms of the arguments given above (§2.1.2 Stuff Happens) by Humphries (1997a): objects are formed by the fusion of parts; no new parts emerge, but a new whole does; it is a whole in that it has causal properties that are not merely the causal properties of the parts added up. The causal trajectory of the parts is tied up with that of the whole it is a part of to the extent that in order to predict this trajectory, you would have to refer to the whole.

Therefore, physicalism can be concluded without ruling out emergence. Emergent causation organises matter in such ways as to have certain causal properties (it builds machines), which are determined by the evolutionary process not the constituent matter (the power to cause things being derived from their being designed to extract, store and use energy from the environment). These are kinds of causation not captured by physical science, which is interested in causal regularities holding between its own fundamental kinds (that is, if physics is interested in causes at all).

This view of objects seems intuitive, which I take to be an advantage, but it does need to be fitted with an appropriately philosophically sophisticated account of causation in general. Such accounts exist, for example in the notion of 'mark transmission' (Salmon, 1984), where a thing can be said to have diachronic continuity to the extent that the stages of that thing are sufficiently causally related. The next chapter is going to explore in detail the concept of causation (§3.1 Causation), advocating a 'difference making' account, linking this to explanation and scientific laws (§3.2 Explanation). We will then be in a position to give a detailed characterisation of notions that have been used liberally in the above, namely supervenience and realisation (§3.3 Supervenience & Realisation), leading to a characterisation of emergence (§3.4 Emergence).

Chapter 3: Levels of Causal Explanation

3.1 Causation

What caused the explosions in London on 7/7/2005? There are many possible partial answers: British foreign policy; the emergence of radical political Islam; evil terrorist criminals; the beliefs & desires of the four bombers; the triggering of an electrical current in the presence of acetone peroxide; the birth of my daughter during the night. Apart from the last one, which would be an irrational belief resulting from an attempt to make sense of the coincidence of two significant but unrelated events, all the others seem to be valid to an extent, in that they seem to have some explanatory value. Explanation will be analysed further below (§3.2 Explanation), but here we will look at the nature of causation and the relationship between the various types of cause.

What makes all but the last causal explanations plausible is that they could be said to have played some important part in bringing about the event, which could be analysed by saying that without being preceded by the aforementioned states of affairs, it wouldn't have happened. For example, in the counterfactual situation where everything else is the same but the terrorists had forgotten to put batteries in their triggering devices, there would have been no explosions. On the other hand, if my daughter had not been born in the night but everything else were the same, we think the explosions would have happened anyway. Another way of putting this is to say that causes are individually necessary and jointly sufficient for their effects. Furthermore, we assume that other states of affairs sufficiently similar to the ones mentioned would have the same effect. For example, whether the batteries are alkaline or lithium based would make no difference, as long as the temperature is held constant within certain bounds. This leads to the important matters of repeatability and generalisation in relation to causation.

I assume one of the main aims of science is to classify the world into kinds with causal powers in order to explain, understand and control the world. I also assume (as a physicalist) that the only way to have causal powers is to be made of (or supervene on) physical stuff. Therefore, in an important sense, our account of causation is prior to our understanding of natural kindhood. If, on our understanding of causation, the only suitable causal *relata* are objects found in the generalisations of basic physics, then there will be no valid causal generalisation involving beliefs *as* beliefs, therefore no genuine causal powers of beliefs, and thus, on the above understanding of natural kindhood, beliefs could not be natural kinds referred to in scientific statements about the causes of actions.

This accords with Fodor: 'causal powers... in the psychological case... supervene on local neural structure.... mind/brain supervenience(/identity) is our only plausible account of how mental states could have the causal powers they do have' (Fodor, 1987, p. 44). Fodor also says that these kinds of assumptions are inherent in the very concept of science itself; that science must operate by classifying things into natural kinds, each of which possesses intrinsic causal powers (Rockwell, 2007, p. 59).

But science does not necessarily require that we can spell out exactly how higher-level causal relations are realised physically. Causal statements are abstracted from individual observations, which generally don't involve the microphysical components of the observed system. Newton and subsequent generations of physicists didn't say how gravity works, but based on observations, we have little problem saying that an unsupported apple will fall to the ground due to gravity. Likewise, the lack of neural 'spelling out' of intentional explanations of action does not invalidate them as causal statements, on the assumption that no 'spooky' action at a distance is invoked.

Thus, an important question for any theory of causation, whether it be a counterfactual analysis, a regularity based account, or one that refers to causal powers, is to what extent it is realist about causation. Are causal statements about person-sized objects merely reflections of our epistemic limitations, which could always be supplanted by the micro-causal story of the composing ultimates, whatever they be? Can we be realist about physical causation without ruling that talk of mental causes be eliminated? (As hinted at, my answer will be that we can be realist about mental causes while holding that everything is wholly composed of physical stuff, which places strict constraints on what things with minds can cause, and what kinds of minds they have.)

Relatedly, philosophers often stress the importance of distinguishing epistemological questions from metaphysical ones: metaphysically, we may hold that all causation is a matter of material micro-bits banging into each other, but when it comes to producing actual statements about how objects will behave causally, we remain limited and so must rely on 'rougher' causal statements that refer to macro-objects and may be riddled with exceptions. However, I question the clarity of this distinction. Do we have any means to answer the question 'What is a cause?' apart from thinking about the world as it presents itself to our scientific enquiries? Can we just assume that the world is such that it is made of atomic particles which do all the real causal work? The argument for physicalism above (§2.2 Physicalism) does not provide sufficient *a priori* grounds for this further claim about causation; it merely says that all causal processes must involve physical parts, that possibilities are constrained by this fact, but not that all such processes are the additive result of the causal processes of the smallest physical parts, whatever they are.

Another problem with relying on the metaphysical hunch that all causal statements must ultimately be 'cashable' in terms of the causal properties of the fundamental things objects are composed of, is that our causal statements are then held hostage to the fortune of causation at that fundamental level. If, as held by neo-Russellians, the notion of cause is inappropriate at the fundamental level, then our higher-level generalisations will suffer 'causal drainage.' If it turns out that the physical ultimates do not have the kinds of causal properties we are used to in the everyday world, for example, if there are uncaused or purely random events, then the tactic of analysing macro-causal statements in terms of micro-causal ones will mean that causal properties leak away through the bottom of your metaphysical bucket. That would leave us with a purely epistemological account of causal statements, and a correspondingly nominalist position on natural kinds. This seems to make the causal generalisations we make in sciences that refer to the everyday objects we see around us, or to mental states, vulnerable to being falsified by empirical discoveries in physics. However, as long as we can construct a robust and physicalistically respectable account of emergence (see 3.4 Emergence), discoveries about the causal nature of base physical reality may explain and constrain the causal nature of larger scale objects without undermining the autonomy of causal statements about those objects.

To put it another way: focusing on micro-causal atoms would mean not seeing the macro-objects they compose (like your body), as the atoms are born in stars and only briefly get caught in the space-time eddy that is you. If objects like you cannot be part of the story that science tells us about the world, and science gives us our ontology, then you do not exist. If you are not so philosophically self-destructive, then you should prefer a view of causation that allows science to include you in the arc of its narrative. To make your body visible to causal narratives, we need a way to draw a boundary, to say when an atom is part of the story we are interested in when we are talking about your body, and when it is not. Such boundaries are not drawn at the level of atoms, even though we may require that the boundary is drawn in a way that is consistent with explanations of higher-level causal properties in terms of lower-level ones.

The problematic assumption that causation is, in the final analysis, a matter of the causal properties of the stuff that composes things, is of a piece with the problematic assumption that causes should be seen as universal regularities (the Humean account). This picture sees the context within which a cause happens as an interference that prevents what would happen in the absence of confounding factors from happening. On these accounts, these contextual factors are 'externalised' (or swept under the rug) from the regularity by *ceteris paribus* (all else being equal) clauses.

However, there are no contextless events, so no 'neat,' universal regularities. In fact, the context enables causation rather than interferes with it. Moreover, if a regularity account tried to include context in a statement of a universal regularity (rather than a tendency as in the capacities account I will be endorsing), the result would be a law for every particular event. These 'laws' would be about never-to-be-repeated states of affairs, and therefore would not function as generalisations to be used in scientific explanation or prediction. All events are particular, but science must try to abstract from particulars to make useful statements; if it tried to cover all particulars without abstraction, this would be like Lewis Carroll's useless 1-1 scale map (see §1.5 Mapping).

The solution, according to Cartwright (1999), is to 'internalise' *ceteris paribus* clauses by 'externalising' causal properties. The Humean approach is to see the causal properties of an object or state as intrinsic to it, as essential, non-dispositional properties, prevented from 'pure' expression by the context. Thus the need to 'cut out' the context in order to see the causal properties express themselves without interference. But nothing happens without a context in which to happen, so this is to assume that there is some sort of naturally neutral background that would allow the cause to bring about its effect unhampered. However, there is no neutral background, no 'view from nowhere.' We can agree on a 'normal' background to take as a base, for example air pressure at sea level on earth in the case of the boiling point of water, but this does not allow us to see the intrinsic causal nature of water at work, since the choice of that background is contingent on facts about the planet we happen to have evolved on.

In the 'impure' real world, where there are always interfering factors (i.e. a context), multiple *ceteris paribus* clauses have to be added to statements of regularity in order to maintain them. Capacity accounts, like Cartwright's, 'internalise' these clauses, since the causal capacity of something is defined as the capacity it has to bring about certain states of affairs *in* certain circumstances. To take a living example, in the case of a micro-organism in the sea that uses light to direct it to oxygenated water, the context (i.e. that light reliably indicates oxygen, due to the fact that water near the surface is more oxygenated) cannot be ignored when describing the causal powers of (parts of) the organism. Change the context, with a torch for example, and the state no longer achieves what it should. A regularity account that ignores external context would only be able to say that the organism tends to react to light, but this does not adequately account for the presence of that part of the animal that has this property. That part of the organism was selected for its tendency to direct the organism towards oxygen in Earthly seas, which it does by directing the organism towards light. This is a more informative, more explanatory (it answers more 'why?' questions) description of the causal situation, and it makes causal properties ones that belong to objects in the world, rather

than being defined by ideally abstract laws that may be expressed in the world in a partial, encumbered fashion. Causal laws are abstract because we do the abstracting for pragmatic reasons, to create useful statements about the structure of the world. They remain statements about real objects in a messy world, rather than Platonic laws that only ever find imperfect expression.

Since in a capacities account causal statements are about the kinds of causal capacities particular kinds of objects have in particular kinds of context, certain types of scientist are mistaken in their belief that they are uncovering universal regularities in the world. Such scientists try to shield the objects they are interested in from the world of 'interfering' causes, and so say something universal about them, by isolating them in laboratory conditions, which Cartwright calls 'nomological machines,' as they are built specifically to produce such laws. However, there is still a context, although a replicable one with clearly defined contextual factors. This further demonstrates the error of the metaphysical assumption that the causal properties of objects are the additive result of the causal regularities of the atoms of which they are composed, which assumes that each atom has its own contribution to make to the whole, based on the laws it is following, rather than being itself governed by the whole it is a part of. This view, while it might not be endorsed by physical scientists who see themselves as discovering the universal laws of nature, may serve as a useful corrective to the kind of 'physics envy' that practitioners in other sciences sometimes suffer from.

As an illustration, think of a physicist falling through the air. He might think that there is a natural speed, terminal velocity, that his body 'wants' to fall at, if it weren't for the interfering factor of air resistance pushing against this. He may calculate his actual speed by visualising this in terms of vector arrows with particular velocities. However, the fact is that his body is travelling at exactly the speed it should, given the context. If we took away the air, the context would be different. But just as causes don't happen without a context, physicists never fall in a vacuum.

But, how do we individuate objects? How do we draw boundaries around portions of the cosmos, the contextual causal properties of which we then discover? As mentioned (§1.2 Are Natural Kinds Found or Made?), this depends on our understanding of it as a causal entity. Therefore, we need an account of causation which is suitable both for the initial proposing of kinds of objects according to their typical causal contributions to their contexts, as well as for the subsequent investigations that will refine our understanding of such things. Of course, wholes are made up of atoms and nothing else, so in what sense are the properties of wholes something other than just the properties of those atoms all added up? We can start to give a general answer to this by looking at quantum physics, chaos theory or other examples of non-linear dependence. The global set up of, for example, the double-slit experiment, affects the pattern of interference exhibited, which is not the additive result

of the behaviour of the parts. Those, like Kim, who argue against higher-level causal properties, it seems to me, are implicitly assuming a classical, Newtonian framework. Even in classical mechanics, however, there are dynamic, chaotic systems, like those found in fluids where the context sets up eddies that lead to emergent phenomena like whirlpools. So, even more narrowly, reductionists seem to be relying on situations akin to those nomological machines Cartwright talks of, where contextual effects are, as far as possible, ruled out.

There is a genuine question for positions like the one advocated here to answer, though: am I not moving from the pragmatics of being able to make useful generalisations to some sort of illegitimate reification? Unlike the neo-Russellians mentioned above, I am not depending on there being no causation at the micro-level to reduce to, partly because it seems clear that the causal properties of the constituting matter play an important role in constraining the higher level. I will attempt to answer this question by outlining my view of object individuation in the next section, with reference to difference-making accounts of causation and Mackie's INUS conditions.

3.1.2 Where to draw the line?

> Nothing can better show the absence of any scientific ground for the distinction between the cause of a phenomenon and its conditions than the capricious manner in which we select from among the conditions that which we choose to denominate the cause. (Mill, 1843, p. 198)

Our nomination of certain aspects of the context preceding an event as a, or the, cause of that event, as opposed to being just part of the context within which the cause operates, is dependent on our explanatory purposes. What to hold constant and what to vary is a decision we make relative to our interests. In order to pick out the causes of an event, we have to be able to refer to the objects or properties involved, which requires separating background conditions from causes. For example, when we say the spark caused the fire, the complete causal context includes the presence of oxygen in the atmosphere (without which the fire would not have happened), but we 'background' such regularly occurring conditions. We can distinguish background conditions from causes by looking at the 'nearby' possible worlds: we say the spark is the cause, the oxygen the background, because the nearby possible worlds contain an earth like ours, with an atmosphere like ours. This particular abstraction ('sparks can cause fires') is therefore of use to us.

As with our drawing of lines between kinds of things, the fact that facts about us are relevant to where lines are drawn does not make those lines subjective in a 'bad' way. That is, it doesn't make those lines not part of the real world: given those facts about us, where those lines will be is an objective fact. Mackie's (1974) account concurs, describing causes as Insufficient and Non-

redundant parts of Unnecessary but Sufficient causes (INUS conditions). Here, we don't need to identify *the* cause, as separate from background conditions. The whole situation (spark plus oxygen plus combustible material) is sufficient to bring about a fire, but some other combination could have brought about the same event. However, the spark is not redundant in an explanation of this particular event, even though it was not sufficient in itself to make it happen. When we choose to describe the spark as the cause of the explosion, this is good for teaching how to make a bomb, but not for a socio-political explanation. So, causal properties are relational, context dependant, and we choose among all the causal factors in play depending on our interests, and one of our interests is to discover useful generalisations we can apply in many situations, so we ascribe causal powers to kinds of object intrinsically because they have that property in many situations (knives cut many kinds of things).

For the purpose of explaining actions in terms of their mental causes, we are interested in the kinds of things that can support the sort of causal mechanisms that can explain the causal link between events seen as cause and effect, that is, we are interested in events and their causal properties. As argued earlier (§2.1.2 Stuff Happens), it is a mistake to say that properties are events: events can be described in a variety of ways depending on our explanatory interests, and causes are properties of events. The causal properties can be described in physical terms, or mental ones, depending on our explanatory interests, without suffering the problem of causal exclusion – see below (§3.1.3 Making a difference).

The picture is of causal properties as relations between event types that happen in and to certain kinds of objects. This does not need to be an instantiation of a universal regularity, just something that is, or would be, useful for predicting and explaining those types of events. We ascribe a causal power to a kind of object or event when we can reliably use that tendency to bring about other events in our predictions and explanations. In the example of an explosion, depending on our interests, we could explain the effect by reference to either the causal power of sparks in the presence of volatile chemical mixtures, or the power of radical ideas in vulnerable minds. In both these cases, the putative cause/effect relationship is not an instantiation of a universal regularity that played out without being confounded by other factors (e.g. moisture or surveillance); it is rather a statement of what tends to happen to certain kinds of things in certain kinds of contexts.

This type of explanation, which refers to contextual tendencies rather than universal regularities or laws, clearly requires an alternative model of explanation from the traditional deductive-nomological one, as this relies on the prior existence of laws in the form of universal regularities. This will be the topic of the next section (§3.2 Explanation), but here I will note that the idea that

causes require laws, as in the idea that to be a cause an event must instantiate a universal regularity, or the governing-law model, e.g. (Armstrong, 1978), where it is the law that drives the behaviour of the object, seems, to me, the wrong way round. This would make instances reliant on properties, but rather, properties are abstracted from instances.

Causation happens between individuals, and laws may be abstracted from many instances of causation if we can group them together in a meaningful way. These laws, by which I mean contextualised generalisations, are not accidentally true local facts, and not instances of universal laws or regularities, because they are explained by the causal tendencies of the kinds of objects mentioned in the generalisations, in the kind of context we are taking as background conditions. When we cite such a law in an explanation or prediction, the explanatory or predictive power derives not from the law itself, but from the context-embedded instances it is abstracted from.

3.1.3 Making a difference

Before moving on from this necessarily brief discussion of causation, I will outline and endorse the analysis of causation as 'difference making' (List & Menzies, 2014). The truth conditions for making a difference are as follows: the presence of F makes a difference to the presence of G in the actual world if and only if it is true in the actual world that (i) $F\Box\rightarrow G$; and (ii) $\sim F\Box\rightarrow\sim G$.

This solves a potential problem with the idea that causal relations hold between instances rather than properties, as outlined above, the problem being that this could lead one to make causal statements that are too 'fine-grained.' To use List and Menzies' example, take several instances of a parrot pecking at crimson coloured spots. One might be tempted to generalise from this to 'Crimson causes pecking.' But observations of crimson pecking only give us the first condition, that crimson is sufficient for pecking to occur. It doesn't satisfy the second condition, that pecking wouldn't occur if the spot were not crimson, because it happens that parrots will peck things of all the various shades of red. So, 'Red causes pecking' is the correct generalisation, which may be realised by instances of crimson. They call this 'proportional causation': 'Satisfaction of these conditions ensures that causes are specific enough for their effects, but no more specific than needed' (List & Menzies, 2014, p. 6). As such, this account of causation is suitable for the kind of contextualised generalisations we are interested in here.

One important implication of this account is that the exclusion principle is false (List & Menzies, 2014, p. 10). List and Menzies (2014, p. 3), state the principle thus:

> If a property F is causally sufficient for a property G, then no distinct property F*, that supervenes on F, causes G (given physical causal closure).

In the example given, red supervenes on its instances, one of these being crimson. Crimson is sufficient for pecking, but it is the case, under this analysis, that red causes pecking, because it is not the case that there is no pecking when there is no crimson. Obviously an account is needed for determining supervenience relationships in order for this to work. This will be addressed below (§3.3 Supervenience & Realisation).

Before looking at the implications of the views outlined for the special science debate, we need to look in some more detail at the concepts of explanation and law.

3.2 Explanation

> 3 statisticians go on a duck shoot. A duck flies overhead, one statistician shoots low, another shoots high, the third shouts "We got him!" – David Foster Wallace

Some, e.g. Pietroski (2000), take the view that whereas causation is a relation between events, explanation is a relation between facts. However, it seems to me that both relations rely on abstraction in a way that makes them difficult to pull apart. Facts, whether particular or general, use kind terms and properties to refer to and describe states of affairs. Events, when related causally, are likewise taken to happen to particulars described using general terms (e.g. red things). The picking out of a particular object or event depends to an extent on our categorical tendencies and explanatory interests. Some ways of doing this may be 'better' than others because they more clearly track the dynamics of the actual world, but there is no single, true way of picking them out that is divorced from us.

Of course, there are useful forms of explanation that are non-causal: the explanatory use of non-causal laws (e.g. nothing travels faster than light), mathematical reasoning, symmetry principles, geometric laws, inter-theoretic relations, renormalization group methods, fact causation, statistical modelling, etc. A detailed analysis of these is beyond the scope of the present work, but I would argue that they all 'piggyback' on implicit causal relations (see Skow (2014) for an argument of this kind). I would say that proponents who overstate the importance of these modes of explanation are in the same position as the third statistician in the joke at the start of this section. As with statistics, these methods are useful, but the reason that they are useful is that they manage to capture some causal facts about the world, even if they don't refer to those facts explicitly.

Explanations are often answers to 'Why?' questions, but they can also be answers to 'How?' questions, and both may be required for a full explanation. For example, in biology, giving the function of a trait explains why it evolved; giving a mechanical description explains how it works. There may be many possible 'hows' for each 'why,' and also, given different contexts, multiple possible 'whys' for each 'how' (e.g. exaptation): neither in isolation will give a full explanation.

As an illustration, take the candiru fish of the Amazon. This fish is reputed to swim up urine streams to its food source. (It turns out this may be a myth, but it can still work as an example, swimming up a stream of thought instead.) In order to achieve its ends, it will need to have a way to sense urine and its direction of flow, and to orient the fish upstream. On identifying parts of the fish's anatomy that seem to have a particular function, we can ask the questions 'What is it for?' and 'How does it work?' It could be that we start with a functional understanding and this leads us to look for the mechanism that achieves that end. Depending on our explanatory interests, the answer to the question 'What kind of thing is it?' may require both. The physical description will not be enough, since another creature could have exactly the same equipment for doing a different job, and since it evolved by a different route for a different purpose, it would be a different kind of thing. The functional description will be insufficient because the same job could be done by a different mechanism, also with a different selective history. It could be the case that swimming up a stream of urine was an ability selected for leading the fish to food, or leading it to a place to hide from predators. The question of how a function is achieved is answered by citing the causally efficacious property that underwrites real explanations, but we don't have to name that property itself in order to have an explanation of why a behaviour occurs. Selection-based laws provide descriptions that allow this indirect explanation, because selection-based processes produce variably realised kinds (Papineau, 2010).

If belief/desire explanations really are explanations, then they should be such that they can be used to form causal generalisations, and thereby be natural kinds. For this to be the case, there must be physical, causally efficacious properties that underwrite these generalisations. This, though, leads us back to the possible problem of the purported intentional-level explanation being excluded by sufficient 'subvenient' causes. Before turning to the issue of supervenience (§3.3 Supervenience & Realisation) we will look at the relation between explanation and laws.

3.2.1 Laws

One reason we need to talk about laws is that we need a theory of the mental that contains real laws, rather than dispositional definitions, in order to create a theory that can explain the necessary

connections between the physiological and the phenomenological (Steels, 2003). As we will see later (§6.2 Consciousness), we need there to be real laws involving the phenomenal in order to explain the impossibility of philosophical zombies. If there is to be a genuine science of the mental, one that includes states like believing, then such states should be causally efficacious, and as a result should be useable in generalisation, categorisation, prediction, etc. That is, we should be able to say things about them that have the form of laws. However, objections are often raised to the effect that mental states cannot be subject to laws, not without losing their essential intentionality. McGinn (1978), for example, says that psychophysical laws can be ruled out because mental states are not natural kinds, this being because mental states fail to display the standard characteristics of natural kinds. However, he only considers everyday usage of mental vocabulary, and generalises this to all mental talk (Cooper, 2005, pp. 70-71).

Davidson (1970) says that although intentional states like beliefs and desires are causes, they are singular rather than lawlike, because they are reasons for action, but do not guarantee the action is performed: we can always decide not to do what we want. Davidson sees the problem of psychological laws as their being not strict enough, in that they are 'infested by *ceteris paribus*' qualifications that can never be discharged. Putting aside the question of whether it is possible to act against one's desires freely, without introducing another desire to so act for some reason, the deeper problem that makes it difficult to see mental states as falling under natural laws is the assumption that laws are universal regularities governing the behaviour of things. As argued above (§3.1 Causation), things don't follow laws, laws don't determine what things do; the way things behave determine the causal generalisations that we abstract from events. If we adopt an account of causation that refers to contextual causal capacities rather than universal regularities, then these clauses can be eliminated, the cost of this being to localise the scope of the laws. Since the causal capacities of organic kinds are relative to the environment in which natural selection worked, this should not be seen as a problem. To extend this story to mental states, a selection story needs to be told for them that makes the mental state emergent and causally efficacious (see §§2.2.1 Causal Closure, 5.2.1 Evolution).

It could be objected at this point that, on a view like this, where causal laws are seen as abstractions from individual events, laws cannot do explanatory work, as they are just re-descriptions of the data. My response is that the explanatory work is done by explicit reference to kinds of objects and their irreducible causal capacities. Moreover, this view provides us with a strong response to views like Kim's, which use principles like causal exclusion and closure to threaten the causal efficacy of mental states *qua* mental states.

The error of assuming a governing-law model, where laws are eternal, unchanging facts that determine the behaviour of every physical particle, leads to the also mistaken idea that all you need to do is add up the influence of these laws on each part that composes a whole in order to deduce the behaviour of that whole (see §2.2 Physicalism). But, if laws instead are seen as abstractions, generalisations based on what we know about the behaviour of material objects, then we have no right to infer the universality of the laws we arrive at. This is a version of Hume's fork, for if everything follows universal law, and we know what they are and the positions of all particles, then the rest follows *a priori*; but if all we have is abstraction from limited observation, then nothing is ever certain. The latter certainly seems more like the status of most science, apart from the most mathematical of them, but, as pointed out (§§2.2 Physicalism 3.1 Causation), they may have no place for causation at all, so we can exclude them from the present discussion. It might be that laws have been different at different times in the history of the universe, that as the behaviour of matter changes depending on its context, so the laws change, since there is no such thing as a contextless bit of stuff. So, as the matter contained within an organism has a different context, with properties that depend on its developmental, evolutionary trajectory, then why assume that the particles contained within are simply following the laws of physics, rather than being affected by biological and cognitive laws?

The assumption that causation requires strict laws is a highly idealised picture taken from physics (and one that may not even be true in that field). Capacity accounts like Cartwright's don't see laws in such strict terms, unless it is specified that the system is artificially isolated in experimental conditions specifically designed to isolate particular causes from their normal contexts (Cartwright, 1999). The way the causal closure of the physical is often portrayed is based on these narrow ideas of causal laws. In reality, systems are never isolated like this, and the regularity account leads us to say that in uninsulated contexts there is a fundamental regularity being interfered with. In the case of higher-level kinds of things like organisms and minds, they have causal capacities given their environment, and with kinds like these it makes no sense to ask how they would behave in isolation; their environment is a necessary part of the description of their capacities.

To summarise: a capacities account of causation and laws, instead of looking for universal regularities that may be confounded by context, identifies the effects that kinds of objects tend to have within their context, thus 'internalising' the *ceteris paribus* clauses by making the contextual effects part of the cause rather than an external factor. Rather than asking what universal laws are being interfered with, it asks what particular kinds of things tend to contribute to their causal contexts. The context is thus part of the nature of a thing, making this a natural way to think of

causation in the case of evolved, embodied agents. Laws don't cause anything, things in the world do, according to their nature and context; laws are derivative (c.f. Lewis' (1973) 'Humean supervenience'). Assuming a universal regularity which may be blocked by certain conditions means the job of science is often to specify these *ceteris paribus* clauses. According to capacity accounts, the task of science is to ask what effects things have a (perhaps imperfect) tendency to produce in certain contexts. Thus, with respect to mental explanations, a science of the mind should investigate the mechanisms and processes involved in producing behaviour and the idea of strictly law-like regularities is dispensable.

Instead of trying to infer from observations what the laws are, we instead infer what causal powers objects have in the kinds of contexts in which they are observed. Of course, the usefulness of generalisations is determined by a balance between generality and applicability. We want to be able to apply scientific generalisations to as many situations as possible, without making them so 'ideal' so as to not apply to any in fact; and we want generalisations to apply to real situations without making them so specific that they are no longer general.

A thing's causal nature is defined by its typical causal contribution, and, in the case of evolved kinds, may require reference to its environment and selective history, since here-and-now physical causes may not be enough to determine what it was 'designed' for. Here-and-now physical causes answer 'how?' questions without answering all the interesting 'why?' questions: evolved kinds come to be for a reason, which explains their causal natures.

To repeat, what is needed for explanation in the case of evolved mechanisms, like the ones we are assuming are behind decision making and action formation in cognitive beings like us, is reference to an evolutionary trajectory: what was it selected for? This functional description provides explanation, and therefore grounds our claims that such mental states can be natural kinds as we can make true, causal, generalizable statements about them. These functional properties supervene on the physical properties of the composing matter without reducing to the physical properties of that matter (they are emergent). It is the defence of this claim that will occupy the next section.

3.3 Supervenience & Realisation

A first pass: A supervenes on B iff something happening to A necessitates something happening to B. For example, a belief X supervenes on brain state Y iff a change in belief (to not-X) cannot happen unless some change also happens in the brain-state. The states upon which the higher-level states (e.g. beliefs) supervene, are said to realise those higher-level states. The question is whether this

sketch of the relation between levels of description holds, and what the consequences of its holding would be.

The notion of supervenience is of central importance to contemporary metaphysics and the reductionism debate, as it is advertised as a way of understanding the relationship between different levels of reality within a monist metaphysics. It is supposed to be a way of understanding how certain properties can be real, in that they are physically realised without requiring any extra, mysterious forces, but are at the same time not reducible to statements about the stuff that realises them. However, this raises the question of whether the fact that a mental state supervenes on certain physical states means that the mental state is 'nothing over and above' the physical states. Can we have mental kinds that are distinct from (i.e. irreducible to) physical ones ontologically, but which are composed without residue of physical kinds? Before interrogating the notion in order to extract answers to these questions, I will first make some more general comments.

There are issues around whether the kind of strong co-variation that supervenience offers is necessary or sufficient for the kind of dependence we want to ground our mental talk, and what this buys us at the explanation counter. Is co-variation sufficient for us to say that a subject S is in mental state X because S's brain is in physical state Y? I would say that co-variance alone, no matter how strong, doesn't get you the kind of dependence relation you need to buttress an explanatory claim; not all, or even most, of the interesting 'why?' questions are answered by a mere supervenience claim.

Accepting a supervenience claim without further explanation of why that arrangement of matter has these mental states is unsatisfactory. Identity claims without this leave the connection between the mental and the physical as contingent rather than necessary (Steels, 2003). The conceivability of philosophical zombies (that there could be an exact physical replica of you that lacks your mental states) is an artefact of a mistaken analytic reductionism about mental terms, i.e. that mental terms are synonymous with the physical terms that describe their realisers. Those aspects of mental states we use as 'reference fixers,' such as functional roles, should be seen as necessary in the same way that some aspects of water (e.g. its density) flow necessarily from its physical composition and its context. (See §6.3.1 The Living Dead)

Token identity (the claim that each instance of a mental state is identical with some physical state) preserves physicalism (McGinn, 1978, p. 211) without the problems associated with type identity (the claim that every mental kind is identical to a physical kind). Type identity doesn't allow for the kind of multiple-realizability we require for mental kinds (e.g. that I can share a belief-type with you

without sharing a physical state-type). However, it is unsatisfying to say that the mental kind is identical to the disjunction of possible realisers; this seems to be a kind of gerrymandering. We need to say more about what kind of physical states are suitable for realising such mental states and why, in order to be able to make useful predictions and explanations.

Kim uses the claim that the multiply-realized properties are disjunctive properties, and that such properties are not projectable, to argue that functional kinds are not scientific kinds. Fodor, in his rebuttal of Kim's critique of his earlier theory about Special Sciences, makes a 'distinction between a multiply based property that is disjunctive, and a multiply based property that is disjunctively realized' (Fodor, 1997, p. 153). Pain is a kind because it is a functional property which is projectable.

One way to argue that reductionism of the mental to the physical will not succeed is to argue against the thesis that the mental supervenes on the physical (what McLaughlin (1984) calls a FIST, that is, 'argument by appeal to a false implied supervenience thesis'). Do non-reductive physicalists need to argue against supervenience of the mental on the physical, or is it enough to show how supervenience is misconstrued by reductionist uses of the relation? I will now turn to the ways in which the term can be construed in general, and in this debate in particular.

There are various ways the supervenience relation can be stated and understood. For example: mental phenomena (Ms) depend on physical phenomena (Ps) in that there cannot be any M events without P events; M supervenes on P if being P indiscernible implies being M indiscernible; M supervenes on P if there can be no change in M without a change in P; the supervenience base P for M is all those P-properties that are jointly sufficient for the M properties (Kim, 1992). It can be taken to be 'weak,' as in being true of objects in a particular world, or 'strong,' as in holding of objects across possible worlds (Haugeland, 1982). It is normally taken to be an asymmetric and conceptual relationship, in that M supervenes on P but not *vice versa*, though whether this is something added to the basic idea of supervenience or inherent in it is open to question. Supervenience itself doesn't imply that M is 'nothing but' P, or that M is 'explained away' when we understand how it depends on P. It just helps to understand the relationship between distinct descriptions of the world. Where does the asymmetry enter, and does this change the nature of the relationship?

As mentioned, the basic idea that all flavours of supervenience share is that if M supervenes on P, then there can be no difference in M without a difference in P. This implies that if two cases are P-identical then they are M-identical. It is not taken to mean that P causes M, but that there is a certain relationship of covariance between P and M. Clearly, there isn't the asymmetry in this relation that most anti-reductionists would want: given a constant temperature, the pressure of a

gas supervenes on its volume, and its volume supervenes its pressure; there can be no change of one without a change in the other. Strictly speaking supervenience can capture the kinds of intra-level dependence of causal relationships (e.g. pressure, temperature, volume), or geometrical properties (e.g. the proportions of the sides of triangles), without saying that one is 'nothing but' the other and is thus 'explained away' (Humphreys, 1997a, pp. S339-S340). To get the kind of asymmetrical dependence consistent with our intuition about the relationship between the mental and the physical (that the mental is dependent on the physical in a way that the physical is not on the mental), we have to introduce a 'and not *vice versa'* clause. This asymmetry is what you would expect in versions of physicalism that want to allow for the possibility that there may be multiple ways for supervenient properties to be instantiated materially, since here facts about the realisation base will determine facts about the realised property, but facts about the realised property will not determine facts about the realisation base (although they might be constrained in interesting ways).

What we want from the supervenience relationship is a way of capturing the relationship between realized and realizing properties and objects. Moreover, we want a way of characterising the kind of multiple-realization we need for a functional analysis of mental phenomena. Mental properties depend on physical properties, but physical properties do not rely on mental properties (in the same way). This is a conceptual relation rather than a causal one in that the subvenient physical properties do not cause the supervenient mental properties; to do so they would have to precede them, whereas they are co-instantiated. Thus, the supervenience base for any particular M is all those P properties that are jointly sufficient for the instantiation of M. For illustration, it can be said that a statue of the president supervenes on the marble it is carved from in that the marble can change (e.g. through the chemical effects of weathering) without a change in the identity of the statue, but the statue cannot be changed (e.g. altering the features so it resembles a goat) without also altering the marble.

In 'everyday' use supervenience means (or meant) something like, 'to follow on from': e.g. "By reason of the cold supervenient winter, I was tyed to the bed" (Hume, A. (1594) in (Kim, 1990)). It was a causal notion, but one where the effect is singular and so unpredictable. In philosophy, it is a specialist term, generally taken to be independent of this vernacular sense (McLaughlin, 2008). I'm not sure the disconnect is so absolute, though. It first came into philosophical use with the rise of emergentism in Britain, where it was used synonymously with 'emergent' (Morgan, 1923; Broad, 1925). Emergent phenomena follow on from, become naturally manifest in, but are not explained by, are in addition to, a certain set of circumstances (although the kind of 'following on from' is taken not to be a causal kind), and, importantly, the supervenience of M on P does not necessarily

entail that P explains M. It may be the case that my feeling of ennui supervenes on parts of the physical world, including my brain, but that doesn't mean that pointing at a splash of colour on a brain-scan will answer the question of why I feel like that. Supervenience is still used in this way by many non-reductive physicalists (Kim, 1990). Having said that, it seems true that our use of the term is not constrained by ordinary use in the way that concepts like 'freedom,' 'cause,' or 'good' may be (McLaughlin, 1995).

The term was put to use by moral philosophy (Hare, 1952), to describe the relation between natural and moral facts, without reducing ethical facts to physical ones. Contrary to those who say that there is no relationship between these two kinds of facts, it was claimed that it is inconceivable that two worlds could be exactly alike with respect to all natural properties while differing in their moral properties (global supervenience); there can be no moral difference without a natural difference, but there can be natural difference without moral difference. This seems to capture a strong intuition. How could it be the case that some possible world is exactly the same in all natural properties as this one but where the moral properties are different? To say otherwise would be to accept that while in this world the invasion of Iraq on false pretexts was wrong, there is a possible world in which everything is the same except that it was right.

In this sense there is much to recommend supervenience when it comes to understanding the relation between generalisations using mental vocabulary and the physical world that underwrites all events, if physicalism is true. As laws are statements about general patterns in the world rather than governing principles, the nomological reduces to the definitional; and metaphysical reduction (in terms of everything being composed of physical matter) is achieved by showing how mental properties can exist in a physical world, without implying analytical reduction (which would be the case only if law-like statements involving non-physical properties could be translated without loss into statements referring solely to physical properties).

The concept was introduced into the literature on the mind-body problem by Donald Davidson: 'there cannot be two events alike in all physical respects but differing in some mental respects' (Davidson, 1970, p. 214). He used it to formulate his theory of the explanation of actions, known as Anomalous Monism, the importance of which to the contemporary debate is such that it is worth briefly outlining it here. Davidson takes it for granted that at least some mental events are causes of physical events, that events related as cause and effect fall under strict laws to that effect, and that there are no strict laws connecting mental events. This last condition is the 'anomalous' part, which follows from the way mental explanations cites reasons, which may be beliefs and desires, and are related to other beliefs and desires in a holistic way not found in the realm of physical causes. The

'monism' part is part of the attempt to dissolve the apparent contradiction contained in the three assumptions, and says that mental events are token-identical with physical events. That is, some events have descriptions that are mental descriptions, and also, by supervenience, they have physical descriptions. It is by virtue of having physical descriptions that they can cause other physical events in a way that can be made into a strict law; and it is by virtue of having mental descriptions that they can be related to other mental events in intentional explanations.

Davidson gives us a picture of the relationship that goes beyond 'mere' covariance (which is insufficient to account for the 'asymmetry intuition' mentioned above), by showing how the mental depends on the physical without reducing to it. However, there is a tension between these two conditions: if M's are too 'strongly' dependent on P's, then reduction threatens. If, like Davidson, we don't want mental explanations to reduce to physical ones, then we will not want to endorse too strong a supervenience claim, despite the fact that it allows you to keep your intuitions about what grounds mental descriptions. We will return to the question of the 'strength' of supervenience later in this section.

Kim claims that the tension between being dependent on but not reducible to results in a failure of arguments against reduction that rely on supervenience. For Kim, for a mental state (M_1) to cause another (M_2), given supervenience, it must cause its realising physical state (P_2). This is so-called downward causation. But, given the causal closure of the physical, P_2 must have a sufficient physical cause, and given causal exclusion, if P_1 is sufficient for P_2, then M_1 is excluded as a cause. It is only by being P_1 that M_1 causes M_2, and M_1's causal power is dependent on P_1's. If there is no causation between Ms, there are no causal laws involving Ms, so Ms are not natural kinds.

However, the argument relies on the premises of causal closure of the physical and exclusion, but as pointed out above (§2.2.1 Causal Closure), these are not perhaps as solid a ground for Kim's conclusions as he seems to assume. Moreover, his argument relies on a causal theory of reference of kind terms, where kinds are baptised by acts of 'original ostension' (Kim, 2005, p. 113), which is to assume that we can pick out the essential causal nature of a thing by naming it, which I have also given reasons to question above (§1.3 Rigidity). Most importantly for our present purposes, though, is the fact that the synchronic & diachronic (local & non-local) interpretations of causal closure are seldom adequately distinguished, and once we loosen these binding principles, there is, I will argue, room for an ontologically robust form of emergence. Now, as there are a variety of supervenience concepts to choose from, we will consider whether there is one that is amenable to anti-reductionist arguments.

The ‘strength’ of a supervenience claim could be on a range from the weakest to the strongest. Weakest-supervenience states that the relationship only holds in this world, whereas strongest-supervenience states that the relationship holds in all possible worlds. Towards the weak end, we could have a supervenience relationship that holds in ‘nearby’ possible worlds, which resemble this one in some important respects but not others, for example it contains the same individuals with the same histories except for one fact that we manipulate in a supervenience thought experiment. Towards the strong end, we could distinguish between nomologically and metaphysically possible worlds. If our supervenience concept is too weak, we won’t be able to hold onto our intuition that there is something explanatory in the relationship between the physical and the mental, as it would not necessarily be the case that two physically identical people in nearby possible worlds are mentally identical. If our supervenience concept is too strong, multiple-realization will in effect be ruled out as there will be nothing left to describe after a physical description is given. Kim-style arguments from supervenience to reduction rely on a strong version, as the weak one does not give him the kind of dependence required to ground mental states in local physical properties.

Weaker versions do carry an explanatory burden, as one would have to say why the relation doesn’t hold under certain changes, but I do not think that burden is unwelcome, as I think we do want to give explanations of why the mental supervenes on some kinds of physical processes in some circumstances, rather than just taking it as a brute fact. So, the interesting question is what strength of supervenience we should plump for. Is there a medium strength version that is rich enough to ground mental states naturalistically in the physical world without being so strong as to lead to reduction?

The distinction between weak and strong supervenience is modal, that is, it depends on whether the relation is a necessary one, and what kind of necessity we take that to be (e.g. logical, nomological, metaphysical) (Kim, 1990). When we say ‘No M-difference without P-difference,’ is this meant to be true given the laws of nature, or the laws of logic, or something else? One way of interrogating intuitions on this point is to consider the possibility of philosophical zombies (Chalmers, 1996): could there be an exact physical duplicate of you that is not a conscious thing? If the mental weakly supervenes on the physical, there is no reason such a being couldn’t exist. As well as there being things that we normally think of as minded not being so, there could also be things with minds that we usually don’t think of as having the right kind of architecture to have them, e.g. a super intelligent shade of the colour blue. This fits with the kind of token identity that requires no further explanations. If the mental strongly supervenes on the physical, such things are impossible: any being with the same physical goings on as you will have, as a matter of necessity, the same mental

goings on as you; and things without the 'right' kind of physical realisation could not have minds. According to Kim (1990), Davidson and Hare are using the weakest version, which is not strong enough for the reasons given.

Another dimension of distinction between supervenience concepts is that between global and local supervenience. McLaughlin gives the following definition of global supervenience:

> A-properties globally supervene on B-properties if and only if for any worlds w1 and w2, if w1 and w2 have exactly the same world-wide pattern of distribution of B-properties, then they have exactly the same world-wide pattern of distribution of A-properties. (McLaughlin, 2008)

He uses this to formulate physicalism, which it may be suited to, but for our purposes this suffers from a similar problem to weak supervenience: it doesn't result in a kind of dependence between states and their realization bases that is informative. It is consistent with there being differences that should be insignificant at the subvenient level that lead to wholesale differences at the supervenient level. For example, it is consistent with it to imagine a very nearby possible world where there is one more atom of hydrogen on Jupiter and no mentality on an earth, despite the fact that everything on earth looks the same (Kim, 1990). In other words, it doesn't imply property covariance, and so can't ground intuitions of dependence.

The opposite, a strongly localist supervenience claim, both spatially and temporally, though, is inconsistent with the kind of multiply-realised, embodied, states we need to account for. It leads to the kind of eliminativism that follows from a reductionist argument. Is there a middle way? Can we have some sort of 'regional supervenience' (Horgan, 1982), a form of individual supervenience with some of the advantages of global supervenience? This kind of supervenience of an object over its parts may be called mereological supervenience. In the case of mental states it is usually assumed that we are talking about the brain as the realisation base, but it may be that less or more than the whole brain is necessary for mental states to be instantiated, and that this may vary with the mental state in question (see §5.3 Externalism).

Prinz (2006, p. 18) argues that in order for (what he calls) 'wide' supervenience to be the case (for experience to not just depend on the brain but also aspects of the environment), then it must be that experience can change while keeping the brain fixed. (Prinz is talking about experience rather than mental states involved in action causation, but the arguments are parallel, and indeed inseparable, as I will argue later in §§5.3.1 Physical Bodies in a Social World & 6.2 Consciousness). However, it could be responded, plausibly, that if a change in the environment is to be constitutive of a change in mental state, then it would have to be the case that it is registered in some way in the

brain. But, this wouldn't show that the part of the world in question is not (partly) constitutive of the (experienced) mental state it is involved in, merely that a change in the brain is necessary also.

We can accept that every experience has a neural correlate, while acknowledging that correlation is neither causation nor constitution, for we could resist the solipsism inherent in internalism: an experience of a cup is only an actual experience of a cup if it is in fact caused by a cup, and so the cup can be said to be part of what constitutes the experience, otherwise it is an experience just like the experience of a cup, an *as if* experience that is derivative of actual experiences. Of course, standard, indirect-realist representationalism disagrees with this. On that account, what we experience is the representation, and there is no difference between a veridical and non-veridical experience, except for their truth values. This is equivalent to me showing you a picture that represents my table as having a cup on it; in the two possible worlds where the only difference is that in one there is a cup on my table and in the other there isn't, the picture is identical in all respects, except in one the situation it represents is actual. But this relies on a false picture of representationalism, one that is vulnerable to homuncular attack. The representation is not between our experience and the world, it *is* our experience of the world, and if there is nothing to come between us and what we experience then we are not trapped in a Cartesian theatre experiencing the world only through projections on a 'screen'; we experience the world directly and therefore there is a real difference between experiences that are of things as they are, and those that are not. (See §4.3 Content)

The disagreement between indirect and direct realist interpretations of representationalism is a boundary dispute: where do we draw the line around what counts as constitutive of a representation? We are faced with the same question when we try to say what counts as a local property of an object or state in the sense of what constitutes its supervenience base. This is undefinable separately from where we draw boundaries between what is internal and external to an object, and deciding this depends on a prior understanding of how to individuate objects, so cannot be used, *pace* Kim, as an argument for reducing said objects to their local, microphysical constituents. To do so would be question begging, as it just assumes that the local, microphysical realisation is all there is to being an object of that kind.

Realisation bases need to be causally sufficient for the realised properties, that is to say, the causal properties of the realised state are explained by the causal properties of its realisation base; the relationship between the two is not contingent. This means that the supervenience base must be all those parts of the world that are jointly sufficient to realise the supervening object, and, if certain kinds of embodied, external cognition are the case, this base will be 'wide,' which undermines the

kind of 'narrowly local' assumption relied on by Kim (see §5.3 Externalism). Noe (2006) makes similar points in his arguments against the idea that there are neural correlates of consciousness, as consciousness emerges from the dynamic interplay of neural and bodily action with the world.

So, supervenience should be understood in a way that is sufficiently broad-minded to allow for the emergence of things and states that are not mere aggregates of physical things and states in that their properties cannot be synchronically reduced to the physical properties of local physical realisers. On the other hand, it should be sufficiently conservative to satisfy our intuitions regarding the kind of explanatory dependence we expect there to exist between mental states and their immediate physical realisations. These two constraints, pulling in opposite directions, result in something similar to the internal realism of Putnam (1981), or the position outlined in Mandik and Clark (2002): it is realist in that it provides a real connection between mental representations and the world out there; it is internal in that the form those representations can take not only depends on and is constrained by the way the world is, but also the situated, dynamic trajectory of the organism doing the representing.

One of the constraints may be that a certain kind of realization base may be a necessary condition for a particular form to evolve. It is conceivable that there is a world where organisms are silicon rather than carbon-based given that these share the property of easily forming compounds: given the kind of flexibility at the molecular level that is necessary to give evolution the variety it requires to work, there probably isn't a world with gold-based organisms. We can imagine the evolutionary history in the carbon and silicon worlds more or less following the same route. In the case of each evolved kind, traits are selected according to the causal powers they have, with reference to the micro-properties that sustain these higher-level properties being generally unnecessary, since they are invisible to selection: it makes no difference to the process of evolution what element forms the basis of organisms, as long as it is suitable, although of course the element that does form that basis may make a difference to what traits are possible. So, if it is possible that human-like brains evolved in both the carbon and silicon worlds, and if it is true that in human-like brains there are such things as beliefs and desires on which cultural evolution works, description of these mental kinds for the purposes of explanation/prediction needs no reference to carbon or silicon. A suitable realisation base is like a near frictionless surface over which the forces of higher-level causes can slide freely.

Supervenience is a conceptual relation that holds between certain properties and the properties of the substance that realises them, but it does not, I have argued, place reductive restrictions on ontologically distinct levels of properties that emerge through a process of evolution. As with causal closure, a supervenience relation that is sufficiently strong to form the basis for reducing higher-

level properties to their local, physical realisation base, is question-begging in that it merely stipulates that the local physical stuff that constitutes the mental state is sufficient to fully characterise that state. Loosening supervenience to allow spatially and temporally extended states allows for 'robust' emergence and downward causation:

> Supervenience is acceptable as a consistency condition on the attribution of concepts, in that if A supervenes on B, you cannot attribute B to an individual and withhold A from it…. But supervenience does not provide any understanding of *ontological* relationships holding between levels. For that emergence is required. (Humphreys, 1997a, p. S341)

Explanations of why and how one kind of property supervenes on another (e.g. functionalist or teleological ones) are grounds for supervenience claims, and without at least the possibility of such explanations, these claims would be ones of faith. When we have such explanations for supervenience, this has been called 'superdupervenience' (Horgan, 1993). Such explanations are a physicalistically respectable way of substantiating claims of emergence, which we now turn to.

3.4 Emergence

Emergence is the key to the metaphysical discussions outlined above. If objects and states with novel properties, novel in not being contained in, or explicable in terms of, the properties of the realization base, can emerge somehow, then there will be true generalisations we can make about them that will not be reducible to statements about the composing matter. If higher-level kinds are emergent against a background of the matter from which they emerge, then there must be something wrong with arguments that purport to show emergence to be impossible. That may seem a strange way to argue, but I think it is a reasonably intuitive starting point.

Strawson (2006) argues that if we accept mental phenomena are real and distinct in kind from the phenomena described by physical science, then either mentality emerges from the physical, or it must have always been there in some form. Given that there are good reasons to reject the latter option (see §6.3 Panpsychism & Composition), then we should accept that emergence is possible. Therefore, there must be something wrong with arguments that purport to prove its impossibility. As pointed out above (§2.2.1 Causal Closure), this may be the use of causal closure principles, the proscription against over-determination, or both, and may result in a breakdown of (certain varieties of) supervenience. In this section, I will distinguish various meanings of emergence, and defend the one I think most relevant for and true of mental phenomena.

Firstly, there is no clear reason why the burden of proof should fall on the defender of emergentism rather than the champions of reduction, unless the question is begged by a metaphysical assumption

about the causal nature of the world, i.e. that all causes are describable by physics, or that physics is complete. But physicalism does not require this; just that there is a natural explanation of how objects and states come to have the causal powers they do, through processes that have happened to things wholly composed of physical stuff without 'outside' intervention from mysterious sources. One reason physicalists seem to think emergentists owe them an explanation is that they think of that which emerges as being non-physical, and indeed some emergentists are out as dualists. But emergentism can be defended within physicalism, and much of the talk of strict distinctions is a hangover from the dualistic past. The distinctions between what is describable in mental terms and what is describable in physical terms is not a strict one; as pointed out above, the relationship between the physical and mental is an informative one. Once the air of mystery about the mental is dissolved, we will see that the emergence of the mental from the physical is a natural thing.

Silberstein (2002) says that both emergentists and reductionists feel that things are going their way, and agrees that the burden of proof is not carried by emergentists alone. In the fight between reductive and non-reductive physicalism, both accept supervenience, but differ in their criteria for identifying natural kinds of one level with another. For reductionists, identification depends on the presence of necessary links between higher-level and lower-level kinds, whereas for non-reductionists it depends on whether or not the higher-level property enters into causal generalisations that do not follow necessarily from the generalisations of the composing material (Silberstein, 2002, p. 104).

The essential ingredients in the recipe for emergence in a physical world are time and some 'looseness,' in the sense of Cantwell Smith's (1996) 'flex and slop.' He uses this phrase to capture the kind of open flexibility that exists in the 'sloppy' physical world, in contrast to an imaginary 'Gear World,' where every atom's motion is intimately connected to its neighbours in a way that allows for no freedom: 'Moths can fly into the night with only a minimal expenditure of energy, because they have to rearrange only a tiny fraction of the world's mass' (Cantwell Smith, 1996). In the Gear World there would not really be any individual entities; everything would be intimately connected in a way that did not allow for distinctions. It takes time, plus a little 'flex and slop,' for the stories of component particles to become so intertwined that they can't be understood without reference to the whole dynamic they are a part of. This 'gappiness' in physical law is what provides the 'elbow room' for entities to emerge over time, allowing dynamic processes of feedback to forge homeostatic wholes.

Humphreys (1997a, pp. S341-342) gave a number of criteria characterising emergence. To summarise the important ones, an emergent property is novel, in that it is a property that had no

instances previously. An emergent property should also be one that cannot occur at the lower level (e.g. liquidity cannot be a property of a single water molecule), and is covered by different laws from its constituents. As in the water example, emergent properties are holistic properties of the whole. Finally, emergent properties are the result of nomologically necessary interactions between their constituent properties.

This final criterion, though, is silent on whether it is a synchronic or diachronic condition. There is a difference between the way liquidity emerges from the interactions of water molecules, and the way a mature organism emerges from the interactions of the carbon and other atoms that constitute it and have been part of its evolution and development. The latter is essentially a dynamic, diachronic process, whereas the former is just what happens when a sufficient quantity of H_2O molecules are collected together under the kind of conditions normally found on the surface of the earth. It is these necessarily diachronic kinds of emergence that I will focus on, and interpreted in this way I will take the following conditions to be necessary and sufficient for such ontological emergence as is found in the case of evolved, minded organisms.

Novelty is not necessary in itself, as it is possible that by coincidence the same property emerges more than once, although it will be novel in the narrow sense of being new relative to the process in question. (This is analogous to the distinction between being psychologically creative and historically creative in Boden (2004).) Thus, one of the conditions is nomological novelty: emergent properties fall under causal generalisations that do not hold of the constituent parts. Secondly is holism: emergent properties hold of wholes, these being collections of parts bound together by being part of a temporally continuous process. Lastly, the emergent properties should be non-mysterious, in that their emergence can be explained by reference to a temporally extended dynamic process.

I have argued for the existence of natural kinds in the domain of intentional cognition (§1.8 Cognitive Kinds), with the understanding that natural kinds enter into irreducible scientific generalisations, where these generalisations capture the causal capacities of situated objects. If this account is right, then nomological emergence, where the laws that hold of higher-level entities are not determined by those that hold of the constituent matter, must be true. Mere mereological emergence, which applies to the properties of wholes, is too weak. This is because it cannot account for the emergence of natural kinds if natural kinds are defined by their shared causal nature, since having a shared causal nature will lead to there being true causal generalisations, i.e. laws in the sense argued for above, about such kinds.

In the classic account of nomological emergentism known as British Emergentism, e.g. Broad (1925), there were still bridge laws between the levels of description, but these laws were taken to be 'brute facts,' with no explanatory connection that would allow determination of the higher level given the lower-level facts. In this account, there is a relation of strict token identity between the levels, and global supervenience is unviolated.

However, this is not the kind of pluralistic nomological emergentism advocated by Cartwright (1999), and Dupré (1993). In these accounts there are no bridge laws; the physical level provides only necessary conditions for the higher-level facts, leading to possible violations of global supervenience, because lower-level laws are not sufficient to fix higher-level laws, although the average effects of the probabilistic causal capacities of the composing stuff may be consistent with a variety of different higher-level capacities. Similarly, McDowell (1994) talks of the lower level as providing necessary but insufficient 'enabling' explanations for the higher level.

I find this account appealing, and amenable to the picture of the inter-level relations I am outlining. The lower level explains the possibility of the higher level, without explaining why the particular possibility that becomes actual does so. To fully explain the properties encountered at the level of evolved creatures and developed minds a further story of the historical trajectory of those kinds of things is needed. Before fleshing out this account, we should first clarify something about the relationship between epistemological and ontological emergence.

Epistemological emergence occurs when the fact that higher-level properties are novel with respect to the lower-level properties is a reflection of our epistemic capacities or limits. For example, in the case of chaotic systems, the properties of wholes cannot be predicted or explained by reference to the properties of the parts, and nor can those properties be fully represented using the theoretical or representational resources sufficient for understanding the parts. The higher-level theory having a pragmatic advantage over a lower-level theory would be sufficient for it to be emergent in this sense. In these cases neither mereological nor nomological supervenience are necessarily violated. This is consistent with the position of ontological reductionism, implying that given a reduction in philosophical, theoretical or empirical ignorance the illusion of emergence could disappear.

This would leave us in the unsatisfying position of believing that the only reason we explain the actions of other minded creatures by referring to their beliefs is because of our own cognitive lacking. But it may be that our higher-level explanatory theories are emergent due to the fact that the laws of the higher-level theory refer to entities that are ontologically emergent in the sense that

the lower-level phenomena (a description of the parts that compose the whole at any one time) are necessary but insufficient for the presence of the higher-level phenomena.

The kind of emergence I take to hold in the case of mental phenomena, in that the causal generalisations we cite in explanations and predictions are emergent in the ontological sense that they refer to phenomena whose causal properties have emerged through feedback dynamics over time. It is not mere epistemological emergence, since the reason those properties cannot be reduced to the causal properties of the composing mater is not just the result of our shortcomings as scientists. However, as argued earlier (§1.2 Are Natural Kinds Found or Made?) it is also the case that the categories we use in our explanations are partly dependent on our perspectives and interests (there are many accurate maps that can describe a territory), and the choice of explanatory schema will depend on pragmatic factors as well as facts about the world. This is consistent with a causal-mechanical model of inter-theoretic reduction that rejects microphysical reduction and 'focuses on explanations as characterising complex (nested and interconnected) causal mechanisms and pathways' (Silberstein, 2002, p. 100), rather than nomological explanation. This allows for multilevel descriptions of causal mechanisms (e.g. genetic, biochemical, intracellular, neuronal, muscle cell, and environmental levels), that fit with the virtual machine functionalism to be defended below (6.1 Virtual Machines).

The idea that there are different laws at different levels leads to a potential problem: the danger of nomological conflict between laws at the emergent and lower level if the former do not reduce to the latter. This problem is only real, however, if the view of causal laws is overly rigid. If laws at all levels are 'gappy,' given time and energy, new dynamic wholes can emerge which follow their own rules, albeit being constrained by the matter they are made of. It is these constraints that avoid the conflict.

A more complex picture is emerging, where 'emergentism and reductionism might form a continuum and not a dichotomy' (Silberstein, 2002, p. 99). With the rejection of synchronic physical causal closure, a time-slice view will fail to capture all the relevant dynamics of entities that have evolved over time within particular environments. Emergent kinds are defined by the composing matter, the environments they have developed in, and the history of the interaction of these forces; a gene for something is only a gene for that thing in the environment in which it was selected as such. Psychological kinds emerge as a result of there being brained beings around negotiating the physical and social environments, and passing on behaviours, beliefs, and so on, to others.

If we broaden our notion of supervenience to include the process of the emergence of a thing and its causal powers, and weaken causal closure to include such diachronic processes, then it can no longer be said that all the causal properties that something has at any point in time is just the 'adding up' of the causal properties of the thing's constituent particles; they themselves are 'pulled along' in this process. The resulting picture is this: a mental state (M) supervenes on physical states ($P_1…P_n$) iff there can be no difference in M without a difference in some P within the transitive causal closure of M (see §2.2.1 Causal Closure). This is a very 'loose' kind of supervenience claim, i.e. not so 'tight' as to rule out downwards causation through the sort of supervenience argument deployed by Kim.

Strawson's (2006) argument against emergence (see §6.3 Panpsychism & Composition) relies on the supposed impossibility of 'radical kind emergence' (Van Gulick, 2001). This kind of emergence is where the properties of the kinds at the lower level do not necessitate the emergence of the higher-level kinds. This is contrasted with everyday emergence, e.g. the liquidity of water, where the emergent property is wholly dependent on the properties of the parts even though the parts don't have this property individually. Water molecules are nothing over and above H_2O molecules, and it is the nature of H_2O molecules that when you put a lot of them together at room temperature on earth, they will slosh about in containers in just the way we expect liquids to. Strawson makes it clear that he's not talking about an epistemological notion of emergence, but an ontological one. His inconceivability argument against the emergence of the experiential from the non-experiential assumes that 'the experiential divide, assuming that it exists at all, is the most fundamental divide in nature' (Strawson, 2006, p. 15) (this despite his criticism of eliminativists for making a similar bold claim; if you assume such a strict divide, of course it will be difficult to bridge). I will give reasons to doubt the strictness of this divide later when we look at experience more closely (§6.2 Consciousness). For now, what Strawson wants to convince us of is that for something so radically different in kind to emerge, something external and mysterious must intervene, something in addition to the matter, which would be a form a dualism instead of emergentism.

Strawson says that for Y to emerge from X, X must somehow be intrinsically suited for Y-type phenomena, and that this intrinsic suitability, in the case of experiential phenomena, must itself be a kind of experiential phenomena, a proto-experiential kind of thing. The former claim is trivial, in that it goes without saying that there must be something about X that makes it the right kind of stuff from which Ys can emerge (given the right context and history), otherwise no Ys would emerge. But that latter claim is non-trivial, and he admits that it is his unargued for intuition that the wholly non-experiential cannot give rise to the experiential, just saying that 'the intuition that the non-

experiential could not emerge from the wholly experiential is exactly parallel and unargued' (Strawson, 2006, p. 21).

My intuition is that these two positions are not parallel, since experience is something that comes about as a result of complex interactions of matter, but I accept that this is question-begging against Strawson's claim that experience could be a simple thing, in that there may be forms of atomic 'proto-experience.' However, if a convincing story can be told of how experience could emerge from complex interactions of matter, involving sense organs, nervous systems, brains, etc., then the existence of such a possibility would be enough to make the two intuitions non-parallel; the burden of argument would then be on the panpsychist. I don't think they do have an argument, aside from the fact that those with the opposite intuition don't have one, so if I give a plausible argument, then I would regard that as putting non-panpsychist physicalism in a strong position.

The argument for panpsychism turns on the notion that the experiential is so radically unlike non-experiential matter that the former cannot plausibly emerge from the latter except mysteriously. However, it is not clear in which respects the experience of a 'full-blown' mind resembles the atomic proto-experiences that supposedly compose it, and without an understanding of in what ways these resemble each other, it is not clear to me how emergence of his conservative kind (from micro-minds to macro-minds) is any more plausible. Strawson concedes that 'human experience or sea snail experience (if any) is an emergent property of structures of ultimates whose individual experientiality no more resembles human or sea snail experientiality than an electron resembles a molecule, a neuron, a brain, or a human being' (Strawson, 2006, p. 27). However, we do understand the relationship between electrons and molecules, and it isn't one of resemblance; we do not say that molecules emerge from electrons anyway. Rather, electrons are parts of molecules, and we don't need to say that there is something 'proto-molecular' about electrons to understand how, put together with other things in the right way, they are part of what makes molecules. If there is no relation of resemblance between the experientiality of the composing matter and that of the organism, it is not clear to me what the argument is, or why we should use the word 'experiential' to describe the properties of stuff that is suitable for putting together such an experiencing system. We may not know what it is like to be a sea snail, but we can make sense of there being something it is like to be one. We don't think there is something it is like to be a grain of sand.

The thing that is special about the matter that composes organisms like us, I will argue, is not that each bit of it is a bit experiential, but rather something about the way it is organised. If you cut out a small part of my brain and put it on the desk in front of me, I will have no problem saying that it itself doesn't experience anything, but when it was in my brain, it was part of a whole system that all

together constituted something which did have experience. Strawson says that every object is a process, but seems to miss the importance of this, which is that certain kinds of process can cause the emergence of phenomena that are wholly novel with respect to the properties of the objects participating in that process. Moreover, processes happen at various levels, with processes at higher levels happening to the objects that are part of those processes, even though, when viewed individually those objects are themselves processes with respect to the stuff they are composed. Obviously, these layers of processes may influence each other in complex and interesting ways, but for the purposes of explanation, it is practical to treat the composing parts as objects with respect to the processes they are part of: although molecules of water are themselves ultimately some kind of process, from the point of view of the process that is a whirlpool, they are objects. Experience is a kind of process that happens to evolved organisms like us (and perhaps to other suitably organised creatures), and viewed in this way, it is not so hard to accept that experiential processes can happen to things composed entirely of non-experiential objects.

Strawson's mistake, in my opinion, is assuming a synchronic and localist supervenience relation. He is insisting on a scenario where indeed any kind of emergence would be mystical. But as argued above, the kind of processes necessary for the emergence of mental properties require time and space to happen. Strawson admits that his answer to the supposed problem of the emergence of the experiential from the non-experiential, i.e. that everything is experiential, faces the serious problem of understanding how all these little experiences add up to make one big experience. We will return to this below (§6.3 Panpsychism & Composition), but now I will return to Kim's arguments, and look again at his arguments against emergentism.

This is Kim's argument against emergentism:

> "...on emergentism, mental properties must have novel causal powers.... [which] must manifest themselves by causing... physical properties or other mental properties. Assume then that mental property M causes another mental property M*.... M* is an emergent; this means that M* is instantiated on a given occasion only because a certain physical property P*, its emergence base, is instantiated on that occasion.... the only viable way of [reconciling M*'s causal dependence on M with its causal dependence on P*] is to suppose that M caused M* by causing its emergence base P*." (Kim, 1992, p. 136) in (Humphreys, 1997a, pp. 2-3)

To paraphrase: since emergent phenomena are necessarily realised by the physical states that they supervene on, then in order for one emergent state (or event/property/object) to cause another, the preceding emergent state has to cause the physical state that realises the later emergent state. He goes on to wield the principle of the causal closure of the physical as a means of severing this putative relationship: since all physical states have sufficient physical causes immediately preceding

them, then, on pain of overdetermination, the preceding emergent cannot be the cause of the physical realisation base of the caused emergent state.

The mistake in the argument, according to Humphreys (1997b, p. 14), is to say the emergent-level phenomena are instantiated 'only because' the emergence base is instantiated. If M* emerges as the result of diachronic processes, where causal closure is also interpreted diachronically (§2.2.1 Causal Closure), both M and P are necessary. It could be argued (R. Chrisley, pers. corr., 2007) that this gets the counterfactuals wrong, because if P* were the case, but not M, M* would still hold. But, if P* is a 'fusion' in the sense introduced earlier (§2.2.1 Causal Closure), i.e. a whole where the parts are no longer seen as independent from the point of view of causal processes the whole is involved in, then P* would not be the case if it weren't for the ongoing process that M describes a part of. That is, we can only pick out the parts of the world that constitute P* by reference to the mereological whole it composes.

I will now briefly outline some of the implications of accepting this view with respect to the issues at hand. Firstly, lower-level theories can be an essential part of the explanation of higher-level ones without being a reduction of them, and the emergent phenomena will thereby not be mere 'brute facts' (Silberstein, 2002, p. 101). Reduction will be at best a local affair, rather than a grand unity of science, with domains forming 'nested hierarchies,' overlapping but not co-extensional. For example, we may be able to give a full explanation of how the mental states of humans have arisen, but that will not be to reduce mental states *tout court* to physical states, as those mental states could also be instantiated in creatures having physical instantiations that are totally different with respect to the properties of the physical parts taken in isolation. Properties in one domain may be necessary but not sufficient for the emergent properties, and the 'local reductions' that can be performed don't violate mereological supervenience, with lower-level entities merging into fusions in a way that means they are no longer best seen as separate entities (Humphreys, 1997b; Boyd, 1991). In such a world, the divisions that are made between levels will depend on the questions being put to nature. This does not mean that the structure of the world depends on us; we discover what the levels are once we have defined what questions we are looking for answers to. This lack of a 'strict' boundary between levels removes the 'air of mystery' around such phenomena (Humphreys, 1997a).

The non-mysteriousness of emergence can be illustrated by the non-trivial examples that exist within the physical level itself. Quantum entanglement is a case where individuals 'fuse' in such a way that they are no longer really individual, in that the 'state of the compound system determines the state of the constituents' rather than, as supervenience normally requires, the other way round

(Humphreys, 1997b, p. 16). Such instances emerge 'horizontally' rather than 'vertically.' Vertical emergence occurs where the higher level is composed of lower-level entities without becoming a dynamic whole; that is without becoming a causal entity that singular causal statements can be made of. For example, a species emerges from the individuals that compose it, but causal generalisations that attach to that kind apply to its instances. Horizontal emergence occurs where the composing parts become a single dynamic whole about which singular causal statements can be made. For example, individual organisms, although composed of parts, are more than the generalised causal properties of those parts, and causal statements about those individuals may be instances of causal generalisations.

Accepting emergence does mean giving up on a certain sort of 'ontological minimalism' (Humphreys, 1997a, p. S337). But, while it is a good thing to embrace pluralism in this messy world, we should retain the explanatory power that comes with minimalism. That is, we should avoid the mistake of trying to produce a map with ratio 1:1, as that would be useless, but if you want to navigate your way over a narrow mountain pass, a 1:250,000 ratio is almost as bad. We want a map that captures in as neat a way as possible all the features on the ground that we need to be able to negotiate. So, ideally, we should have as few ontological commitments as we can get away with, while making enough distinctions to be as accurate as we require. The implication is that there may be multiple valid ways of describing the world in terms of the kinds of things it contains, but that each of these is 'tethered' to reality because of the existence of a causal story that says how those features were formed, and what 'shape' they are. Unlike Dennett's stances, this allows real explanation, rather than mere predictive utility.

Another non-mysterious case of emergence is that of self-maintainant systems, which are systems that can maintain themselves in a far-from-equilibrium state through using resources from the surroundings, e.g. a candle pulling in fresh oxygen to fuel itself (Bickhard, 2006). Organisms are such systems, and have the further advantage of flexibility, being able to change what they do to maintain themselves in the face of a changing environment, like sensing and moving in the direction of nutrients. When explaining the actions of such organisms, we can refer to the functions of those parts of the organism that reliably take in information and produce motion. In such cases, where there is dynamic interaction with the environment to maintain a non-equilibrium state, there should be a way for the organism to process information about the environment, i.e. representational content. Representations, in this view, are not states of some homuncular viewer of inputs, but rather 'future oriented models of interactive anticipation' (Bickhard, 2006).

This brings us to things like us. Here, I am going to sketch the landscape that I will spend the second half of this text filling in the features of. We are in contact with a world the features of which we differentiate on the basis of presuppositions based on the outcomes of previous interactions. These presuppositions take the form of expectations which may be more or less true or false. This dynamic feedback-based process, which is at the heart of multiple levels of selection (natural, social, developmental, etc.), sets up the conditions for emergence. Representation is an act, an event in the process that is our being in the world, an interaction rather than a passive processing of inputs. The view of representations given below (§4.2 Representations) is one that is explanatorily grounded in the physical mechanisms of the brain and body in the world (a world we have shaped in order to ease our interactions with), thus avoiding radical scepticism, and the frame problem (see §4.1). Scepticism is neutralised because the representations interact with the world in a way that means that errors are detectable by the system in principle. The frame problem is avoided due to the implicitness of the 'dynamic presuppositions' in play when we are anticipating our interactions with our world (Bickhard, 2006). In the following, I will be putting flesh on these metaphysical bones by incorporating recent work in cognitive science in a way that will result in a coherent philosophical narrative of how such things as minds come to be in the natural world, a story that will retain the characteristics of having minds that we take to be important, that is, that we can be agents of our own destiny.

Chapter 4: Kinds of Mental Content

4.1 Mental Kinds

Having found our metaphysical bearings, let us now turn towards our objective: the ontological status of those kinds of things referred to in the generalisations that underwrite, implicitly or explicitly, intentional explanations of action. There are a diverse range of states that could be called mental that are involved in the production of action: phenomenal, affective, intentional, to name a few. Therefore, as a starting point, we will start with the broad 'folk philosophical' definition: action producing mental states are any that can be referred to in true explanations of actions, such as propositional attitudes of the form ('X believes/desires that Y', where X is an agent and Y a proposition), or affective attitudes ('X did Y because she was in emotional state Z').

If necessary, finer distinctions will be made as we progress, but as a starting point this liberal, almost trivial definition will suffice. A science does not have to define its terms strictly from the outset; we can start with 'folk' uses and proceed by means of 'interactive conceptual refinement' (Sloman & Chrisley, 2003, p. 143), where neither our theoretical notions nor the empirical data are given a veto over the other, but rather both are in a 'constructive dialectical opposition.' As such, we can take the folk notions of belief and desire in explanations of action and see how far they get us. Our vague cluster concepts of kinds of matter were refined when we learnt about the 'architecture' of matter, and the same might happen with our concepts of mental states (Sloman & Chrisley, 2003). Taking an 'architectural' view, looking at how functional parts are organised into a whole mechanism that exhibits the kinds of properties we are interested in, allows us to investigate a space of possible ways of building systems like that, rather than attempting to define what mind or consciousness is from the outset (Sloman & Chrisley, 2003, p. 163).

We can start with the observation that the states we are interested in are states of minded entities that enter into explanations of action, and add that a condition of being minded could be that they enjoy non-epiphenomenal phenomenological aspects (see §6.2 Consciousness). Third-person descriptions of mental activity may overlap with more typically first-person descriptions in some instances, for example in describing the affordances given in perceiving the world (Gibson, 1979), these being a necessary part of action explanations. It can be posited that the states in question are 'discrete, semantically-evaluable, causally-effective states, possessing component structure, and where those structures bear systematic relations to the structures of other, related, thoughts' (Carruthers, 2002, p. 658). These thoughts may be like bits of language, or could be like images, or models, or other representational structures, and the type of structure they have may influence the type of effects they have. We might want to make a strong claim about the role of language in

cognitive processes; it may have more than an input role in development and enculturation, furnishing the mind with concepts, it also could have implications for building the foundations of the cognitive system.

Before continuing, we need to deal with the idea that the phenomenal properties of experience, the 'what it feels like,' are epiphenomenal. In many respects, this is the same discussion as the one above (§3.2.1 Laws) about the causal efficacy of mental states *qua* mental states. There, we were discussing different ways of portioning the physical world in third-person terms, and asking whether macro-states composed of smaller parts could be ascribed causal properties separately from the causal properties of those parts. Here we are discussing whether phenomenal properties correlated with physical states can be said to have causal properties separately from those physically defined correlates. The argument goes like this: since all phenomenal feelings are accompanied by physical goings on, and the physical goings on are sufficient for the immediately following goings on, there is no causal work for these phenomenal properties to do, given that phenomenal properties are distinct from physical ones. If, in order to avoid redundancy, phenomenal experiences are not identified with some physical state, i.e. dualism is accepted, then intractable questions about how the non-physical can interact with the physical could be raised.

Given that these experiential properties seem to be such an integral part of my mental processes when consciously making a decision to act, it is worth asking, as with emergentism, why the burden of proof is generally seen to be on those who argue for the efficacy of such phenomena. Again, much of the debate seems to be a hangover from the debate between dualists and materialists. But you don't have to be a dualist to argue against epiphenomenalism; we can see experience, I will argue, as a part of the real world that emerges due to physical processes (see Chapter 6: Physically Embodied Minds). Part of the problem with taking phenomenal properties to be a genuine part of the world of causally efficacious things, is certain types of properties often associated with phenomenal experience, like incorrigibility, ineffability, privacy, etc. For the moment, I ask the reader to suspend these assumptions. Later (§6.3 Panpsychism & Composition) we will see some reasons to doubt that these are in fact necessary properties of first-person experience. Here I will briefly cast doubt on one of the most popular arguments for epiphenomenalism.

Arguments such as the knowledge argument (Jackson, 1982) seem to beg the question. This argument relies on the premise that Mary knows everything physical there is to know about experience of the colour red, but learns something new on experiencing the colour for the first time. This is to assume that you can have all the physical knowledge without having seen red, an assumption that relies on a deeper assumption regarding the nature of knowledge, that is, a

'discursive assumption,' that all knowledge can be discursively communicated (Chrisley, 2010). If there are forms of knowledge that are abilities (knowledge-how rather than knowledge-that (Ryle, 1946)), it could be the case that knowing what it is like to see red is a kind of ability, and therefore a kind of knowledge that is not discursively transferable (just as you can't tell someone how to ride a bike). So, putting question-begging assumptions about the nature of the material and the mental aside, there is no reason that the burden of proof must be on the phenomenal realist; quite the opposite since epiphenomenalism is counterintuitive; we certainly seem to experience thought processes as the causes of actions.

Now, given that the reduction/emergence debate as I have framed it revolves around the question of the existence of natural kinds, the question is whether mental states, perhaps with partly experiential descriptions, count as such. Can they be intersubjectively referred to and are they projectable? I will say yes, given the preceding accounts of the relevant terms. If Sloman and Chrisley (2003) are right to say that, as some of the data that enters our explanations of mental phenomena are qualitative rather than quantitative, we need an explanatory theory more concerned with abilities than predictions and correlations, then the mark of a good theory will be that it makes sense of the abilities we find that humans have, rather than it being determinate enough to ground quantitative predictions, or its pointing to a particular physical state, the properties of which will be discursively communicable, that underwrites mental states. I agree with this in spirit, and think Cartwright's (1999) capacity account of causation (§3.1 Causation) is suited for the job, foregrounding as it does the causal tendencies of kinds of things (objects, states, events or properties thereof, including phenomenal ones) in certain contexts.

McGinn (1978, pp. 196-7) offers a Cartesian argument from Kripke against the possibility of natural kinds in the mental realm, based on the conceivability of physical realisers and phenomenal states 'coming apart.' It is an *a priori* argument regarding the necessity of identity. In the case of *a posteriori* identities like 'water=H_2O,' water is baptised ostensively, then its essence discovered empirically, with the connection between facts about the physical essence and observable facts about water being metaphysically necessary (once we know all the facts about hydrogen and oxygen and chemistry and physics, etc., all the facts about wateriness follow). In the case of phenomenal states like being hungry, so it is argued, there is no physical essence to pick out ostensively. Even if we find the physical correlates of such feelings, there is no metaphysically necessary connection that would allow one to infer one from the other. The same holds of the functionalist arguments to variable realisation. That means we cannot point to a paradigmatic example of a proposed mental kind, define an equivalence relation for other members of that kind, and provide or discover an

acceptably empirical 'real essence' of that kind (McGinn, 1978, p. 197). So, we know *a priori* that mental states are not members of the natural kinds (McGinn, 1978, p. 199).

My first response to this kind of objection is that it begs the question in one of two ways. Either he is assuming that empirically respectable properties have to be described purely in the language of physics, that 'real essence' has to be couched in terms of the properties physics talks about, or he is assuming there is no physical essence to pick out. In the latter case, it could just be a familiar case of our referential abilities outstripping our understanding. In the former, he needs an argument for why states like being in pain cannot be defined functionally without reference to physical correlates and still be empirically respectable properties capable of projection.

McGinn (1978, p. 202) gives more reasons why functional or dispositional properties are not suitable candidates for natural kindhood. First, because of the holistic nature of mental states picked out by propositional attitude statements, each one will not have a causal profile unique to it. Holism, indeterminacy and the rationality constraint separate mental ascriptions from physical ones. Mental states like these mediate between perceptual inputs and behavioural outputs and are individuated conceptually by their place within a conceptual scheme, where each item depends on others within the system for its definition. Thus, mental kinds do not fit with a causal theory of reference but a descriptive one, and so are not natural kinds. Similarly, 'real essence' must be a property of 'internal structure,' but mental states don't have one. The dispositional properties of mental states like beliefs and desires are their essences and will not be reduced to physical states; that would be to change the subject (McGinn, 1978, p. 201). Thirdly, specification of causal roles is '*a priori* and definitional,' unlike descriptions of 'real essences,' and if essences are not discovered *a posteriori* then they will be nominal rather than real. Lastly, mental states are attributed on the basis of behaviour, but unlike other ways of attributing natural kinds to objects, these cannot be later refined and replaced by more essential descriptions (McGinn, 1978, p. 214).

But, as we have seen, the causal theory of reference may also be inadequate for non-mental kinds, as a descriptive element is necessary for referential fixing there too, so this argument is undermined. Either we accept that almost all natural kind ascriptions include contextual, holistic, mereological descriptions, or we say that there are no natural kinds save for the fundamental things out of which everything is composed. The debate is in danger of becoming a semantic rather than a substantive one here, but, if it is scientifically desirable for water and tigers to be natural kinds, which it should be on the assumption that our aim is to make generalised statements about things in the world, then we should be expansive rather than restrictive, and accept the criteria for natural kinds given above (

Chapter 1: Natural Kinds). Furthermore, it is not the case that our definitions of mental kinds in cognitive science are immune to revision due to empirical input, if the 'Cartesian' insistence on absolute incorrigibility is weakened: the first person, together with normative constraints of the concepts we use, and our physical realisation, jointly determine the structure of our mind, and each of these can influence the others.

There are mental states which have properties that are most properly picked out using lexical concepts, therefore such states are subject to the holism that is a part of the meanings of terms in a language, but they can nevertheless form the basis of predictions and explanations. They rely on a social world because they wouldn't be the states they are if the subject hadn't been enculturated in the linguistic community that he in fact was. For example, in Turkish culture, among others, there is a particular concept of family honour that is connected to the behaviour of the family's womenfolk. If a female member of one's family has a relationship with a male outside marriage, then the men in her family will be in a certain state (*namussuzlukta* – 'being in the state of lacking honour'), which has, for those experiencing it, a particular phenomenal character, and will lead them to behave, with a level of probability, in certain culturally defined ways. The same set of events would have different consequences in other cultures.

The fact that we ascribe mental kinds on the basis of behaviour, whether that be verbal reportings of first-person experience or observed non-linguistic behaviour that reflects mental states of being, leads to the problem of distinguishing genuine cases of acting for reasons, and other cases of acting *as if* there were a reason, that is, between metaphorical and veridical uses of intentional explanation. If we can find no hard and fast criteria for the distinction, should we dissolve it? Perhaps the intentional level of explanation is a fiction constructed retrospectively to rationalise events that are 'merely' causal.

I think this would be throwing out the intentional baby with the dualistic bathwater. We should be able to distinguish usages like, 'The water droplet wants to hit the ground' from, 'I want to hit the ground.' There may be interesting border disputes in the case of simpler organisms (is an explanation like, 'The ant followed the pheromone trail because it wanted to find the food source' metaphorical or not?), but these discussions about where to draw lines are informative rather than problematic. The thing that makes an explanation of my actions intentional is that there is a state that has the function of motivating action; it has that function partly because of its causal profile, but also because it has been designed for such a purpose by evolution; that is, it is a mechanism for which we can give explanations of how it works, how it got there, and what it's for.

What about, 'The missile wants to hit the ground'? There may be cases, in the future perhaps, where this would be non-metaphorical, and to convince people of this, we would need to give similar kinds of explanations as we give in our case. One thing that most people would say is missing in the case of missiles is the phenomenological aspect, which in our case, we are assuming, is part of what makes an intentional state the state that it is. Those aspects are what it's like for me to want something, and, putting epiphenomenalism aside, being like this is part of the state's contribution to the causal situation. If it is the case that phenomenal states play a causal role *qua* phenomenal state, then philosophical zombies are not possible; they only seem conceivable because of hidden assumptions about the natures of the physical and the experiential. Rather, phenomenology is a necessary part of certain states that emerge in creatures like us, and there is no in principle reason that such states shouldn't emerge in machines under the right circumstances (see §6.3.5 Selfishness). In fact, they should emerge necessarily, not contingently; it is the intuition that the relationship between physical and phenomenological is contingent (because they are so clearly distinct in kind) that is responsible for the supposed conceivability of philosophical zombies.

On the other hand, mental states have more than their phenomenological aspects as part of their identity conditions. Philosophical behaviourists like Wittgenstein and Ryle argued against taking the meaning of words that refer to mental states as picking out essentially first-person facts. Much of the vocabulary used to denote mental states develops, evolutionarily, culturally, developmentally, in a complex interaction between reflections on inner states and observations of the states of others. When we say, 'X is thinking,' this seems to imply both internal and external facts. The external aspect might not seem obvious, but we wouldn't ascribe 'thinking' to someone who seems to be engrossed in a physical activity that takes complete focused, in-the-moment concentration, like climbing the crux of an exposed route (although the minutes of indecision before committing to the move could definitely be called thinking). The combination of causal and descriptive theories of reference gives a picture something like this: we learn to apply behavioural descriptions to others when learning a language, then we learn definitions and descriptions which we apply to ourselves, and subsequently we may learn some detailed science about the brain and cognitive systems in general. This may change slightly the concept we originally learnt when some behaviour was pointed out to us, just as our concepts of 'whale' or 'water' might have changed slightly since they were originally pointed at and named. In both cases, part of that refinement is a conceptual endeavour, and part of it empirical.

Furthermore, intentional concepts like 'belief' are not purely first-person, only applicable in our own case with certainty, but depend also on dispositions to behave in certain ways in certain situations.

Because they don't rely solely on their first-person content to be the states they are, then it is not necessary to be conscious of being in a particular state, or even be so disposed to be so, in order for that concept to be correctly attributed. In the case of ascribing motivational mental states to non-linguistic creatures, we are assuming that they have mental states sufficiently similar to ours if we ascribe, non-metaphorically, mental states to them. It is not just a matter of whether we can successfully predict action, but whether that action is brought about by mental states that deserve to be called mental states because of their functional role in a mental architecture that has been designed to house such states. Again, it is the mistake of taking the first-person aspect of intentional concepts as being wholly private, etc. that leads to philosophical confusions like the conceivability of zombies (Steels, 2003). As Chrisley & Sloman (2016, in press) have argued, there may be first-person, indexical concepts that are only applicable in one's own case, but which are nevertheless public. For example, I have the concept of myself, which only I can use of myself, but I am public. The same might be true of first-person concepts of qualia that are only applicable to oneself, but which could be ruled on by third persons.

Davidson (1984) argues that ascribing mental states like the ones we use in explanations of our behaviour to animals is a mistake. He gives the example, 'The dog believes there is a cat up the tree.' He says that this could only be used to explain the rationality of the dog's action if the dog were a language user, since by substituting co-referential terms we can make another proposition that would explain the behaviour just as well, e.g. 'The dog believes there is a cat up the oldest visible tree.' In other words, the contents of beliefs are more fine-grained than what they refer to, so, we cannot be justified in ascribing particular belief contents to animals on the basis of behaviour alone. Davidson (1985) adds that being rational, and thus having your actions explicable by reference to propositional attitudes, requires being a user of a system of propositional attitudes, understanding what it is for a belief to be true, being able to give reasons that one believes to be true, and therefore being a language user. In the case of non-linguistic creatures, then, using intentional explanations is, according to Davidson, metaphorical.

I think this requirement, that being a language user is necessary in order to have beliefs that are reasons for actions and can therefore be used as parts of explanations for those actions, is too strong. There is good reason to think that dogs possess the concepts CAT and TREE, since both are recurring parts of the environment that a dog needs to be able to navigate. We could test for this in various ways, like the ability to re-identify, cross-classify, etc. However, here Davidson has a point, as we could never be fully justified in these ascriptions because such behaviour would never be sufficient to allow us to distinguish between co-extensional but conceptually distinct content.

Obviously, the dog's concept of 'tree' will not be the same as ours, which includes, for example, how to distinguish bushes, but then again, most humans get by without holding in mind the fully-fledged, latest version of the scientific concept of tree. Non-scientific humans and dogs have their intensional content that they use to act in the world with greater or lesser success, and the ascriptions we make of them, or rather, the way we characterise their beliefs in terms of concepts, will predict their behaviour with greater or lesser success (Bermudez, 2003), the difference being that we can use intentional content (*de re* as opposed to *de dicto*) to judge the truth value of the contents of human beliefs, because we have a publicly accepted measure in terms of the meanings of words.

This doesn't mean that dogs don't have beliefs about the world that can be reasons for actions, just that it is problematic to ascribe the content of their beliefs in terms of a *that* clause. But there may be such cases of non-conceptual belief contents that explain actions, including in animals like us, and in those cases, it will not be possible to have thoughts about those contents in the way that is required when stating the contents of a belief in a *that* clause. Still, it makes sense to say that a dog reliably tracks cats, and can track one to a tree, and can therefore be said to believe that there is a cat up that tree, because it would be surprised after chasing the cat up a tree to discover that the cat had disappeared. What a dog can't do is *know* that it believes there is a cat up a tree; for that, language is required, so that the contents of the belief can be stated in a publicly assessable manner that the creature in question could take a reflexive attitude towards (c.f. (Malcolm, 1972 -1973)).

What beliefs and desires may be attributed to an organism is not purely a matter of observer interpretation; such attributions are limited, and made true, by 'architectural' constraints, that is, by the various evolved mechanisms or states that serve the non-trivial needs of the organism (Sloman, et al., 2003). For example, fish probably don't need an overbearing sense of ennui; it could serve no function in their lives, and to interpret its behaviour as such would be anthropomorphism. This is in harmony with Lloyd Morgan's canon: 'In no case is an animal activity to be interpreted in terms of higher psychological processes if it can be fairly interpreted in terms of processes which stand lower in the scale of psychological evolution and development' (Morgan, 1894, p. 59). As well as behavioural evidence, it is not inconceivable that we could bring neurological evidence to bear. We could investigate the internal mechanisms of dogs and discover neural correlates for tracking cats and objects like trees. Then, when we say that 'the dog believes there is a cat up the tree,' we are not saying that the dog believes the statement 'there is a cat up the tree,' rather, the dog believes the state of affairs referred to by that proposition is the case.

The types of action producing states possible are constrained by the structures of the world and the creature. Rather than talking about beliefs and desires, which may mislead in bringing to mind the explicit tokening of propositional thoughts, we can instead talk of belief-like and desire-like states (Sloman, et al., 2003), which are mental states that perform the traditional functions of beliefs and desires respectively. Desire-like states function to motivate acting on and in the world in a way that serves the needs of the creature (Sloman, et al., 2003, p. 19), for example 'preferences, pleasures, pains, evaluations, attitudes, goals, intentions, and mood' (Sloman, et al., 2003, p. 20). Belief-like states function to provide information about the world which can be acted on according to what is desired, for example 'beliefs (particular and general), percepts, and sensory information states' (Sloman, et al., 2003, p. 20). Such states have semantic content in so far as they place truth conditions on the world: they can be wrong by misrepresenting the world.

Some may be purely reactive, some simple 'what if' deliberative systems, some reflective 'meta-management' systems (Sloman, et al., 2003, p. 29). Each 'architecture' makes possible certain states, which may coincide with or refine our pre-theoretical, 'folk' concepts. We can further distinguish reactive (e.g. startle), deliberative (e.g. worry) and reflective (e.g. self-disgust) levels of affect (Sloman, et al., 2003, p. 15). Each will have its own evolutionary history that explains its functioning, for example, more purely reactive ones like alarm signals will tend to be older ones. The higher cognitive functions of being able to deliberate about mental states explicitly, as we do when we explain our actions propositionally, requires the ability to chunk the world into discrete classes and to learn generalisations about them in order to predict, plan, and explain (Sloman, et al., 2003, p. 28).

This sketch forms a picture of mental states that, I will argue, are not subject to the worries of reductionism, elimination, or epiphenomenalism given above. Mental states emerge from the interactions of parts of physically realised virtual machines (see §6.1 Virtual Machines) and environments, rather than being discreet, atomic entities obeying their own laws, or being composed of such objects. Minds emerge from brains in bodies in the world over time, but, although every mind we know is neurally implemented, minds are by definition implementation neutral, as they are defined in terms of the abstract architecture of the virtual machines (Sloman, et al., 2003, p. 40).

The fact that much of the vocabulary used to denote mental states develops, evolutionarily, culturally, developmentally, in a complex interaction between reflection on inner states and observation of others, leads to another objection to counting kinds of mental states as natural kinds: our classificatory practices can affect human kinds in ways they cannot affect natural kinds. Hacking

(1990) says cultural feedback (being classified, or not, a certain way causes people to alter their behaviour), and conceptual feedback (the availability of new descriptions makes new kinds of behaviour possible) makes human kinds subjective. For example, if I am told I have ADHD, and I accept that diagnosis, this will affect my behaviour in a way likely to cause that diagnosis to be a self-fulfilling prophecy. Hacking (1990) takes being a kind to boil down to being law governed, and further claims that our classificatory practices can affect human kinds in ways they cannot affect natural kinds. The kind of feedback involved in self-classification means there can be no observer-independent laws to be taken as the essence of the kind. In the case of classifying human kinds, being classified a certain way can have moral overtones (e.g. 'pervert,' 'normal'), as well as institutional benefits or costs (e.g. 'obese,' 'pious'). This leads to people behaving in ways so as to change how they are classified (e.g. 'person with stapled stomach,' 'woman who wears the veil').

However, it is also the case that non-human kinds like species are affected by our classificatory practices. Cooper (2004) gives the example of dogs, which we would generally say is a natural kind, but which nevertheless has been altered by us through selective breeding and development. Hacking replies that the feedback in human kinds is different because it occurs as a result of subjects becoming 'aware of the ways in which they are being described and judged' (Hacking, 1997, p. 15, as cited in Cooper, 2004, p. 78).

I agree with Cooper that although the case of human self-ascription is different in interesting ways from other kinds of classification, this difference is not fundamental in a way that rules out human cognitive kinds as being natural. This is just part of the causal history of these kinds, and as such will be part of the information that makes them useful for forming generalisations and making predictions. Hacking argues that the idea-dependence of human kinds makes them subjective, but, as pointed out above, all sorts of natural kinds involve some variety of idea-dependency, without making them useless for the purposes we require them for. Cooper (2004) distinguishes two different ways things can be idea-dependent: relational and causal. An example of the former would be that our judgements about the aesthetic properties of certain pictures might change due to changing ideals of beauty. Causal idea-dependence is exemplified in cases like changes in ideals of beauty causing a woman to go on a diet, or a man to moisturise. This is an objective effect on things in the world, and so is compatible with objective judgements of kind membership (e.g. the kind of man who moisturises).

Hacking claims that the human kinds that result from such feedback from descriptions to behaviour are not natural kinds, as they are logically dependent on the existence of those descriptions. However, although it may be the case that culturally evolved descriptions can have a causal effect on

behaviour, it is not the case that the ability to act in a certain way is logically dependent on that description. Cooper criticises Hacking for following Anscombe's (1957) account of intentional action: 'An action X can be said to be intentional when the actor could respond to the question "Why are you doing X?" by giving a reason for acting' (Cooper, 2004, p. 80). She says that if the agent cannot answer the question, then, on this account, the behaviour would not be intentional, which would make being able to answer the 'Why?' question a necessary condition for an action being intentional. But this leads to the conclusion that a pre-linguistic human could not intend to do anything. However, we could say that 'Ug intended his banging the rocks together *qua* a way of making a fire,' using inference to the best explanation (Cooper, 2004, p. 82). Some actions may indeed be logically dependent on the existence of descriptions, for example taking an oath, but most are not. Though those actions may be causally dependant on those descriptions, they are not made possible by the existence of those descriptions: you can still act like a man that moisturises without that description being available to explain your behaviour.

So, an agent need not apply an intentional description to her action for that action to fall under an intentional description. However, I want to sound a note of caution about the priority of certain descriptions in the explanatory pecking order. In the case that an agent is in an explicit intentional state, then, I would argue, in the absence of interference, the intentional level of description 'trumps' other descriptions in terms of being the best explanation of the behaviour; that is, the intentional explanation is irreducible and more informative. Moreover, even when the intention is not conscious, or not even disposed to be consciously available, the intentional description may be best.

However, despite this irreducibility of the intentional, it is a mistake to 'build' from the top down in the way many realists about belief/desire explanations do, for example in Fodor's Language of Thought thesis. The information-bearing states that play the role of beliefs are not like full-blown beliefs in many ways, for example they may not conform fully to the rules of compositionality, systematicity, etc., like linguistic tokens do. Such not-quite explicit states may not be 'inferentially promiscuous' (Hurley, 2003), but that doesn't prevent them from playing the role of beliefs in intentional explanations. In fact, the case of fully conceptualised beliefs may be quite rare, even in the case when we do ascribe the holding of beliefs to ourselves.

One advantage of seeing the conceptualised self as emergent from the more normal pre-conceptual, yet still intentional, self is that there will not be a strict and mysterious discontinuity between us and non-linguistic animals or pre-linguistic infants, and that ascribing beliefs and desires to such creatures in order to explain their behaviour will not always be metaphorical. This is not to deny

that there is, in the case of certain organisms at certain times, an important and distinct kind of reflective self-awareness that requires the development of a self-concept, an 'I', and the constructing of narratives that that self plays a role in. But the often precarious existence of reflective selfhood should not distract us from the reality of pre-reflective selfhood which exists in the everyday coping with being in the world as a creature that navigates that world through sensory experience (c.f. Butterworth (1995)). The conceptualised self emerges from, or floats on top of (and often sinks back into) that fluid 'just-being.' An advantage of not taking the conceptualised-self to be the only true form of being a mind is that some problems that dog the computational functionalist, like the frame problem, become less problematic. The frame problem is the problem of framing what parts of a conceptual scheme to update given some new information; to maintain consistency every concept in the scheme would need checking against the new input every time, putting too high a computational load on the system to be realistic. In a mental model where full conceptuality is a special case, then the chain reaction underlying this combinatorial explosion doesn't have the fuel to get going. The challenge is to explain how it is that concept use can emerge from the nonconceptual in a way that doesn't just reduce the conceptual self to the nonconceptual parts that it emerges from (which was the purpose of the metaphysical arguments in the first half of the present work).

Nevertheless, it is not just trivially true that explicit intentional states, i.e. the ones we can report ourselves, must always be 'couchable' in a public language; such states rely on language in a more interesting and important sense, as it is through a feedback relationship with externally existent public languages, and by implication the socio-cultural world, that they emerge developmentally. To investigate the properties of these states, then, it may not be enough to look at what has been biologically inherited; we need to also look at cultural inheritance. Understanding the nature of mental kinds means seeing how they result from physical, biological, cultural, social and personal processes, which means not just looking at what is in the head (*pace* Fodor's (1980) methodological solipsism). We may need to look in the head to explain cases where things go wrong, but when looking at mental mechanisms, what they are for and how they are built are both important, and not independent (see §§5.2.1 Evolution & 6.1 Virtual Machines). As well as broadening our view of where to look from a third-person perspective, we need to expand, rather than reduce or even eliminate, the first-person concept of the mental, such that we understand that the nature of experiences may not be fully revealed in experience (Steels, 2003).

Martin (2007) writes that naïve realism (the claim that the kind of experience we have depends on whether or not it is veridical) is inconsistent with two further, commonly held assumptions, namely,

'experiential naturalism' (that our experiences are part of the causal natural order) and the 'common kind assumption' (that the same mental event happens in the case of veridical and hallucinatory experience). Naïve realism would say that there is a difference between the experience of a distant star, and the experience of an extremely large and distant bright thing that is not a star. But, if we assume that mentally we are the same in the case of hallucinating a star and seeing an actual star, and that since our experiences are part of the causal nature of the world there is not enough time for any direct causal influence from very distant objects, then the experiences of the star and the extremely large and distant bright thing that is not a star must be the same. Therefore these assumptions are inconsistent and we must reject one of them.

Before making what seems to be the obvious move of rejecting naïve realism, we should ask, 'What is meant by 'kind' in the common kind assumption?' Martin (2007) says that events cannot be placed under different kinds just because we tend to describe them differently, but that there needs to be some kind of 'privileged' description, which, I will assume, is a description that matches the causal regularities of the real world. It seems clear that, if experiences are physical events (given experiential naturalism), naïve realism directly contradicts the common kind assumption. However, naïve realism as stated is indeed naïve, and is not necessarily the best formulation of direct realism, because it, naïvely, assumes that our experiences are immediately available, or given, to us. But what if there can be a difference in experience that we are not aware of? This may seem to be contradictory on most understandings of what it is to experience something, but later (§6.2 Consciousness) we will look at cases (e.g. blindsight) that make us question this Cartesian picture of experience. If it is the case that there can be differences of experience without experience of difference, then, in the case of experiences of distant stars and not so distant large objects, the experiences *will* be different, even though we can't tell them apart and they appear to have the same causal consequences. However, the fact that we can't distinguish them might be a function of the fact that we haven't probed them enough yet. On that note, in the next section we are going to probe the nature of representations more deeply.

4.2 Representations

Whether the mental states we are talking about are to be thought of as essentially representational is a contentious issue. Some just assume that such states, states that carry information about the world to be used as a basis for behaviour, are representations, but recently there have been strong arguments put forward by anti-representationalists. However, I will contend that the view of representationalism attacked by the antis is a mostly straw man. The view presented here contains

representations, but doesn't interpret this word in the wrong sense as their being the objects of direct experience that indirectly *re*-present the world to a subject, as such interpretations suffer from homuncular objections. Rather, I will use the term in the right sense, of there being internal states whose function it is to carry information about the world for use in certain cognitive processes, for example, when thinking about aspects of the world 'offline,' or when acting fast and fluidly, where there is not enough time for collection and processing of detailed information from the world itself.

For a historical origin of representationalist thinking we can cite Descartes and Locke. Although they may have been on opposite sides of the epistemological debate between rationalism and empiricism, they shared the picture of perception as being a matter of the world being represented to a subject through intermediate internal states between the sense organs and the perceiving subject. For Descartes, this is clear, as the subject is immaterial and receives input from the body. For Locke, the world is presented to the mind via the qualities of experience. Generally, Locke is taken as the source of the position known as indirect realism: roughly, the external world causes impingements on sensory organs forming representations, and we experience those representations. The advantage of such a view is that it accounts for errors in perception: if we directly perceived the world, how could we be wrong about it? The disadvantage is that we cannot be sure we ever perceive things how they in fact are: a 'veil' is drawn between us and the world. The position has an intuitiveness to it: to say we perceive the world indirectly via our sense organs seems trivially true. But if you ask people to look at a patch of red and point to where the redness is located, most people will not point at their head (unless they have studied philosophy). If the redness is in the object, that would be a case of direct perception. However, in the case of phenomena like afterimages, it seems to make sense to say the colour is in the head. The disagreement comes down to whether we take a red afterimage to be of the same kind as an experience of an actual patch of red in the external world.

Arguments against indirect realism have taken many forms, including forms of the private language argument. This argues that it is incoherent to say things like '*This* is what I mean by "red",' referring to an internal, private representation I am experiencing, because the meaning of the word 'red' is a public thing, and must be definable by reference to publically available properties. In other words, ordinarily we use the language item 'red' to refer to a property of things in the world, not in the head. A similar attack directed more specifically at intentional action is Ryle's regress (Ryle, 1949, pp. 30-31). Against the common-sense representationalist picture where there are intentional states with representational content (beliefs and desires) preceding every action, Ryle says that if

having an idea (like believing and desiring a state of affairs) can also be seen as a kind of action, which is plausible, then this too must be preceded by ideas that explain that action, and so on.

Ultimately, such attacks fail as they target a caricature. We don't need a private language to refer to our internal states, just a personal understanding of a public language. Indeed, the redness is in the object perceived to be red, but there is, nevertheless, an internal state in me that performs the function of representing red things, not by standing between my experience and the object, but by just being my experience of the colour of the object. To address the regress argument, not every action is preceded by, explicit, conceptualised representational states, just a special subset, and having a belief or desire is not necessarily in that subset. Some of this has already been covered, and some will be further clarified below, but there are other contemporary attacks to consider first.

There has been a move in cognitive science in recent years towards a more direct engagement with the world, one that doesn't require intermediaries to represent the world to the agent, with much of the cognitive load placed in the world (enactive, externalist, embodied, embedded theories, e.g. (Thompson, 2008)). Much of this line of thinking accords with the present account, applying as it does the principles of parsimony and economy, which are both epistemologically and evolutionarily good: it makes sense to 'use the world as its own representation' (Brooks, 1991) when possible. Why spend resources storing and manipulating complex representations of the world, when the world is right there with the information at hand? There are two questions for such views: does this make them wholly non-representational, and are there not circumstances where it does make sense to use internally represented information as a basis for action?

Many cognitive scientists, e.g. Fodor (1983), have treated the mind as modular, with peripheral systems (perception, action), and central systems (thinking and planning), each 'encapsulated,' using a 'proprietory' code, basically being multiple black boxes taking inputs and giving outputs. Recently this has been questioned, with some (e.g. Prinz (2006)) claiming that the same representations may be used in various mental processes (e.g. both perceptual and motor representations may be used in planning), while others opt for a more radical colonisation of the mind by a particular type, e.g. Noë (2004), who claims that the mental processes involved in action are used also for perception. O'Regan and Noë (2001) ask how representations, traditionally construed, help. For them, seeing is a way of acting rather than a way of representing. The outside world is its own representation: 'seeing occurs when the organism masters what we call the governing laws of sensorimotor contingency' (O'Regan & Noë, 2001, p. 939). This helps explain phenomena like expectation effects, change blindness, etc.: 'vision is a mode of exploration of the world that is mediated by knowledge, on the

part of the perceiver,…. the *structure of the rule*s governing the sensory changes produced by various motor actions' (O'Regan & Noë, 2001, pp. 940-1).

Despite their arguing against representationalist accounts of perception, they do talk about 'cortical representations' (O'Regan & Noë, 2001, p. 968), and the 'lawful' way they change as we navigate the world. In other words, there are mental states (cortically realised) that carry information about the world that we use as a basis for action, and which is based on our understanding of the regularities of the world. So, setting aside the idea of a perceiving subject sitting in the head being presented with detailed, image-like representations of the world outside, planning actions and sending instructions to the body, it is not clear to me how this is significantly different from the 'right' version of representationalism outlined in the previous section. Just because we give up on the idea of a detailed internal representation of the world given to us by our eyes, which acts as the direct object of perception, as an intermediary between us and the world 'out there,' doesn't mean that we should say there are no representations at all, that is, no internal states that are coupled to aspects of the external world and that act as stimuli for our actions. We may use the world as its own representation when that is the most efficient way of doing things, but it is not always.

In fact, I think there is a problem in this way of talking. The world doesn't represent itself, it just presents itself. The representations are the mental states that carry that information. It may be, for example, that we don't represent the positions of all the pieces on a chess board and deliberate on the position internally, the chess board is right there keeping track of itself. But when we pick up a piece we use some internal states to coordinate hand and eye, even if this is done by constant feedback from the world itself rather than being an action plotted according to detailed inner models of the chess board. Interestingly, though, a master chess player can have a detailed inner representation of the chess board on which he bases his deliberations, as shown by the ability to play blindfold chess. Normal people too use constructed models to predict how the world will act in situations where we can't wait for feedback. When acting fast and fluently, rather than performing a movement and getting feedback, which takes time, we predict, or model, the expected outcome of an action based on experience, and feed this information forward. The result of such actions informs future similar actions, either by readjusting the models due to failure, or reinforcing it through success (e.g. Basso (2006)). In short, enactivism has a point: *some* of our way of representing the world may be based on possibilities for action; they present the world in ways that afford certain behaviours, which is based on the models of the world we build up through experience of interacting with it. But this is still a kind of representation, and it doesn't account for all our ways of acting in the world.

As pointed out, anti-representationalists often take representationalists to be saying that mental representations are intermediate and detailed. However, phenomena like change-blindness (Simons & Levin, 1997) (where subjects do not see even large-scale changes in a visual scene) tell against the simple idea that our representations convey the world to us such a way that the details are contained in our representations, waiting for us to pay attention to them, as they show that it can be surprisingly difficult to notice changes that should be obvious; it can be quite difficult to locate significant changes in a visual scene, even when we are looking for them. We can take this to show that we don't have detailed, conscious representations, or that, although the detail is there ready to be accessed, that access is not a simple matter of just looking. Looked at this way, change blindness may not show that we do not have detailed representations, just that we are not aware of changes because we don't consciously compare all details from one moment to the next: 'change of representation is not representation of change' (Chrisley, pers. comm., 2009). Much of the misunderstanding goes back to Locke's idea that there is a relationship of resemblance between the representation and the represented, this being what makes the representation about what it is about. But rather than resemblance, what we should focus on is how the information in a representation is used. An internal representation, or model, need not resemble what it is about in any way. The important thing is whether actions based on it are appropriate in the context (Holland & Goodman, 2003, p. 79).

Neither are representations intermediate between the subjects and objects of experience; having a representation that is used as a basis for conscious action in the world is just what it is to be a perceiving subject; we could call this representationalist direct realism. Moreover, we use this form of representationalism to distinguish things like us, whose actions can be given true explanations in terms of mental states, from things that may merely be metaphorically described as acting for reasons, through the existence of such complex internal states that stand for aspects of the complex world and which are used in the formation of actions.

4.3 Content

Earlier, I advocated a topographical account of natural kinds, which, I claimed, has the advantage of being realist while giving due consideration to how our understanding of the structure of the world develops scientifically, and being able to cope with a messy, uncertain world where what makes a thing the thing it is depends on more than a pure, internal and eternal essence. As such, that account adverts to the structure of both the mind and the world, and this is doubly so when it comes to talk of mental kinds. As scientists describing cognitive kinds, we depend on a mixture of internal

and external facts (*pace* methodological solipsism); as individuals the experience of tokening a cognitive kind depends on more than the brain alone (*pace* solipsism). In both cases, a key feature of a mental state that plays a role in determining behaviour will be the information carried by that state, that is, what it tells the organism about the world, or its content. It is to this that we now turn.

Another way of saying that mental states represent is to say that they have content; they 'say' something about the world in a way that can be judged as accurate or not. For example, a subject may have a belief with the content, 'There is a red snake in that log pile.' This may or may not be the case, but that informational content may be used to explain the subject's reaction. What is the nature of that content? It could be that the vehicles of that content in the mental process between perception to action is a mental image that resembles a red snake, just as a photo of a red snake can be said to represent a snake by virtue of resemblance (Locke, 1700). Another possibility is that the information is carried in the way that sentences in a language do, with symbols that have semantic content and syntactic structure (Fodor, 1975). A theory of the motivational mental states behind action needs to give an account of the nature of the content of these states and how it gets to be there.

There is an important, if controversial, distinction to be made between conceptual and non-conceptual content. Roughly, if the cognisor possesses the concept that is used to pick out the mental state, it is conceptual. Possessing a concept means having some mental content that satisfies the generality constraint (Evans, 1982), which requires the ability to re-identify objects picked out by that concept, and classify together objects that are similar enough to each other to fall under it. For example, an adult human will normally possess the concepts RED and SNAKE, which are tokened when they perceive a red snake, meaning they are able to group together other objects that fall under those concepts. In addition, they may be able to infer that an object that falls under both those concepts is likely to be venomous, and so should be avoided. On the other hand, a pre-linguistic infant will not have these concepts (assuming they are not innate), so will lack the classificatory abilities of adults, and also the fearful reaction to potentially venomous creatures. The mental state that represents the colour of the thing in the log pile to the infant does carry information about that object, but that information is not correctly named by the linguistic concept RED. If we as observers were to name that non-conceptual content, we might want to give it a label that allows us to identify it, but it would have to be something less general than RED, something that is 'narrower,' in that it refers not to the broad category 'red,' but to the specific wavelengths that excite that particular activity in the visual system.

I will for present purposes purposely ignore the disagreement about whether there is such a thing as content that is not conceptual and take both conceptual and non-conceptual content as necessary for a science of mental states that can talk about both language wielding and non-linguistic creatures, as well as the non-linguistic action of linguistic creatures. In fact, the distinction is likely a continuum rather than a dichotomy, and most of our mental lives are spent '…awash in a partially differentiated, partially objective, mind-world continuum…. The exercise of a concept is the result, both literally and metaphorically, of our ability to find our way in the environment; to stay afloat' (Cussins, 1990, p. 411).

A related distinction is that between narrow and wide content, that is, between content that involves referring to kinds of things and properties in the world, and content that is not involved in reference fixing and instead only carries information about the perceptual presentation of experiences (Cussins, 1990, p. 379). To use the example of the concept ELM, you can make true statements about elms, e.g. 'Brighton & Hove has the last surviving population of native elms in the UK,' despite not possessing the ability to identify an elm (the content of the statement does not depend on the contents of your head). But your actions in relation to elm trees seem to rely more on your internal resources: in order to say, 'That is an elm' while pointing at a tree and be right (ignoring the possibility of being right by chance), you need to be able to token a mental state with the appropriate broad, conceptual content.

This is distinct from the question of whether the supervenience base of psychological kinds contained in intentional explanations is 'broad,' which is a causal rather than an informational matter (see §5.3 Externalism). That is to say, the question of whether our concept of content is broad is metaphysical, whereas the question of whether the laws of psychology are broad is empirical (Fodor, 1994, p. 24). A potential problem arises here, in the form of 'twin' cases: on Twin Earth, where 'water' refers to XYZ, watery thoughts leading to actions involving the clear liquid lying in puddles will have the same narrow content (ignoring the complication that we might know something about chemistry), but different broad content, and so will be involved in different causal generalisations, assuming those generalisations are stated 'broadly.' If psychological laws refer to intentional, broad contents, but their implementations are narrow, Fodor thinks there must be a mechanism that causes covariance between the two in such a way as to make twin cases seem strange (i.e. not very common) (Fodor, 1994, p. 27).

The solution may lie in non-conceptual content. The sameness of the narrow content in twin cases depends on not knowing about the chemical composition of water, and this lack means, in effect, that the concept of water is not fully realised, as not knowing this will lead to failures of

generalisations, as the twin case highlights. If we assume full possession of the concepts in both worlds, i.e. a mental representation that contains sufficient information to avoid failures of generalisation, then the narrow content would not be the same, and we should distinguish the concepts tokened in watery thoughts in this and that world. However, if we insist on full concept possession in this sense, then our psychological laws will be restricted to a small set of actions of a small set of creatures.

Fodor's answer relies on the relationship between subjects, propositions, and the world. In the real-world example of the concepts ELM and BEECH, informational semantics equates content to the dispositions of application (Fodor, 1994, p. 30), so a subject who can't distinguish between two different natural kinds does not possess the concepts of those kinds. But, the application of the concepts ELM and BEECH have different truth conditions, therefore differing in content, leading Fodor to conclude that a 'broad' psychology fails to see why subjects' behavioural dispositions should be identical in both cases (Fodor, 1994, p. 34). He states, semantics is about counterfactual dispositions and shouldn't be confused with epistemological questions of what one needs to do to check if one's thoughts are true (Fodor, 1994, p. 37): if one needs to distinguish between an elm and a beech one will consult a book; most of the time, it doesn't matter. Broad, intentional psychology can be true despite such Frege cases if such cases are not 'systematic,' that is, if they are exceptions that can be handled with *ceteris paribus* clauses (Fodor, 1994, p. 46). Nor do we need to fall back on narrow contents in the exceptional cases, as propositional attitudes are *three* place relations between a creature, a proposition, and a mode of presentation (Fodor, 1994, p. 47). The content of the proposition is broad, but the mode of presentation is different, which explains the different desires (and different behaviour) that result. The fact that an individual's behaviour is caused by his or her narrow, internal, processes doesn't mean that psychological laws and explanations should refer to narrow content, but rather, psychological laws are broad and supervene on individuals, each of whom may have a slightly different mode of presentation of the broad content. Fodor concludes that as long as we are reliable devices for arriving at beliefs and desires that permit us to act rationally in the world we are in, then those modes of presentation will succeed in their roles, and intentional laws will supervene on individual minds (Fodor, 1994, pp. 51-4).

This view certainly has its merits, but a major problem with it is that it relies on there being an attitude taken towards a mental token with propositional structure, thereby excluding cases where no proposition-like state is tokened in the mind of the actor, as in non-linguistic creatures and pre-linguistic humans, and indeed normal adults when they are acting fluently and without conscious deliberation. In such cases we might rely on internalised models of features of the world that are

particularly useful, e.g. something like a map, or a list of cues to be followed. Unless we want to take Fodor's route and posit the existence of proposition-like structures in the minds of non-language users, we have to have a more liberal understanding of the kinds of mental content that can enter our psychological generalisations. To see how this might be possible, we should first review the various accounts of mental content used in scientific explanations to see if any of them more naturally allow a place for such content.

The classical theories of conceptual content are imagism (e.g. Locke (1700)) and definitionism (e.g. Russell (1905)), the former being where possessing a concept means having an image of the thing the concept is of tokened in the mind, the latter being where it would mean knowing the definition that uniquely picks out the thing or things the concept is about. Imagism relies on the similarity of the mental token and the thing it represents, and comes in at least two forms, these being prototype (Rosch, 1999) and exemplar (Smith, 1999) theories. In one, when we perceive something, we compare it to stored prototypes, which are sort of average representatives of things, and classify it according to its similarity to one of those prototypes. In the other, we store a set of previously experienced exemplars of kinds of things and compare a new stimulus to those to see if it fits. The problem with this is that it relies too heavily on perceptual similarity rather than something more fundamental, and so can't accommodate examples where, despite a lack of resemblance, we may want to classify two things as of the same kind for other reasons.

Definitionism doesn't have this problem, as the definition will only include how things look if this is indeed essential to being that kind of thing. The problem is that there are few examples of unproblematic definitions that succeed in picking out all and only the type of thing they are supposed to. Furthermore, if we have the wrong definition in mind when we use a word, we will fail to refer to the thing we think the word names. A refinement of this is the theory theory (Gopnik & Meltzoff, 1997), where concepts get their meanings due to being part of a theory about things, a theory that we build like scientists making hypotheses and testing them against experience. This has the advantage of not requiring a specification of necessary and sufficient conditions of application for each term, as the defining happens due to the relations that term has to others in the theoretical structure. The problem with this is that since each individual constructs their own theory throughout their life, and terms are holistically defined by the whole theoretical structure, it is problematic to say that individuals share concepts with each other.

Fodor's proposed therapy for these problems for theories of content is atomism, in which a mental state gets its content due purely to its syntax, that is, the way it relates to other concepts in the system of concepts, rather than being decomposable into semantic parts. It is a causal notion,

where symbols tokened in the brain have the functional role of representing aspects of the world in mental processes, and fulfil this if their causal profile is suitable for it to combine with other items in the mental language in the right way. Being purely formal, this fits with classical computationalism, where the brain is seen as a computer that operates on mental symbols, with the body providing as input and output channels. One problem with this view is an implausible amount of assumed innate conceptual content, as the analysis of propositions into a combination of concepts, and concepts into other more basic concepts, has to bottom out somewhere (Clark, 1994). Fodor seems to accept this conclusion, even in his later work (Fodor, 2008), for much the same reason that Plato (*Meno*, 81c-85d) thought that learning is actually the eternal soul remembering what it already knew: concepts cannot be learnt without already having the ability to token them. There is also the problem of the place of understanding, which is a general problem for any sort of functional role semantics if Searle's Chinese room objection has teeth (see §6.1 Virtual Machines).

All these theories are 'full-bodied' with regard to content, that is, a concept token in the mind contains within itself all the resources necessary to determine what it is about. This fits with Fodor's methodological solipsism: the world provides input, but cognitive scientific explanations need only refer to goings on within the brain. In opposition to this are embodied accounts, which do not rely on self-contained symbolic representations (e.g. Barsalou (2009)). These are such that the role of the body and the world cannot simply be 'bracketed.' Narrow conceptual contents are kept to an economic minimum and much cognitive work can be shared by other, embodied cognitive systems, or off-loaded onto the evolutionary and social niches we develop in.

The result of such a shift in emphasis is a reasonably radical reversal in the role of understanding. Rather than being almost a side-effect of the syntactic, and therefore causal, properties of symbolic representations, understanding is key in the meaningful actions that play a role in concept formation and use. Thus, implementation of mental processes cannot be internalised to purely narrow content; embedded and embodied theories build 'representations' up from physically and socially situated practices, and the syntactic structure of those states can be seen as deriving from, rather determining, what they are about, i.e. their semantics.

In the case of those contentful mental states involved in deliberate acting, and assuming that we to a greater or lesser extent succeed in understanding other folk's actions in order to predict and explain them well enough, when asked, we may advert to beliefs, desires and suchlike in our explanations, even though, most of the time, when just acting, we seem to simply read intentions directly from behaviour without explicitly reasoning about inferred mental states. There are two interconnected questions about this general picture: What are the contributions of innate mental

structures on the one hand, and developmentally inculcated cultural resources on the other, to the cognitive abilities we develop? Also: When we use intentional concepts like beliefs and desires in explanation, what kinds of processes/events must be the case for those explanations to be true? There are two responses to these questions we should be cautious of.

First, we shouldn't assume that the only truth makers of intentional explanations must be discrete, fully conceptual counterparts to lexical concepts contained in brains. As mentioned, there also needs to be room for other states that are more like abstracted models or maps, constructed with non-conceptual content which can nevertheless be called belief-like or desire-like due to the role they play in the system. Second, we shouldn't assume that since such counterparts are not always needed for intentional explanations, then there are no such states in any cases. It is a mistake to look for a single style of explanation; sometimes our behaviour is explicable without positing the tokening of conceptual mental contents, and sometimes the best explanation will posit such things. The world is a messy place and our theorising should reflect that.

Part of the reason for the mess may be that when referring to a theory of mind, we could be referring to a scientific theory of minds in general, or to the mental module that some scientists posit as the evolved mechanism individuals use to understand the behaviour of others. Whether the 'theory of mind' is an innate module (Baron-Cohen, 1995), or the result of a developmental process of hypothesis testing (Gopnik & Meltzoff, 1997), whether we understand others through simulating their behaviour and understanding it *as if* it were ours, or indeed if we do all of these things, such abilities are founded on more basic, embodied (emotional, sensory-motor, perceptual and nonconceptual) practices. As such, the 'second-person' (as opposed to the third-person of theoretical approaches) remains our basic mode of understanding others. Such understanding is 'direct and pragmatic' (Gallagher, 2001, p. 86) because the intention is expressed explicitly in the behaviour rather than inferred from it (as it would be in the third-person mode), but that doesn't mean there are no explicit, conceptualised intentions 'behind' actions in any case. In other words, it is a mistake to take a certain, advanced capacity and inflate its relevance, assuming that it must be like that 'all the way down.' Equally, but oppositely, it would be a mistake to say that because we 'read' intentions directly from behaviour, then there are no states we can call beliefs and desires causing that behaviour: rather, it is a pragmatic point.

Avoiding those errors, I would advocate a middle way. We are not fully abstracted thinkers, deliberating dispassionately, uninfluenced by biology; neither are we purely reactive creatures of the flesh. There are at least two interacting dynamics together forging mental states that can be called intentional: bottom up (non-conceptual, emotional, perceptual, sensory-motor, etc.) and top-down

(theoretical, conceptual, etc.). Both these take place in an evolved system and niche, and include an assumption of a perspective, a first-person viewpoint.

Therefore, I accept Cussins' (1990) statement that a classically computationalist account whose causal connections rely on the syntactic properties of semantically evaluable states 'is unworkable because it cannot capture, in a theoretically adequate way, the cognitive significance of indexical and demonstrative contents' (Cussins, 1990, p. 412). That is, there are explanatorily relevant first-person facts about the way the world is presented to a subject that are not characterisable as the subject's possession of a concept, because the subject lacks the dispositions necessary, for example being able to recognise another instance as being of the same kind.

However, I wouldn't accept Cussins claim that, although 'concept possession would be causally legitimated by the scientific levels of a [non-conceptualist] framework…, conceptual characterizations of content would play no role in the scientific psychological explanations of the behaviour of the system' (Cussins, 1990, p. 437). This may indeed be the case in non-linguistic creatures: we can say that a dog has a concept-like representation of bones (not the human concept BONE) because of the kinds of behaviour towards bones that it exhibits (and its behaviour when there are no bones present), without saying that it is manipulating a symbolic representation with the content 'bone' according to the syntactic properties of said symbol. But a significant difference is introduced in the case of creatures that explicitly use linguistic concepts, where we learn the meanings of linguistic symbols, as well as how those symbols combine with others, during the process of maturation as a member of a language using community. This difference may not be sharp. Words are means of pulling oneself up from the reactive to the more reflective, a gradual process and one that may require some effort, but one whose benefits, in terms of the understanding and the capacities that result, can be passed on to others.

Concepts are used to refer to what the contents of minds tell the creature about the world, or what it knows. The 'gold standard' is knowing that you know, but that is a high bar, and many other kinds of cognition count as knowledge. 'Know-how' (Ryle, 1946), like being able to ride a bike, is of course a special case. I may forget that I know how to ride a bike, for some reason, which will affect my actions when asked if I would like to go on a ride, but that doesn't mean I don't know how to ride one. In terms of 'knowledge-that,' I may also know that bikes need oiling weekly if you live near the coast, but not act on it because I have not had the fact brought to the 'front' of my mind. When someone looks at my rusty chain and reminds me, I may say, 'I knew that'.

So, knowing something is evidenced by being disposed towards certain behaviours in particular circumstances, but is explained by there being states with particular contents that may be triggered by circumstances. As always there are caveats: how much prompting is allowed before we say someone doesn't know something? Thus, if a prompt is needed to remind me of something ('Isn't it a special day tomorrow…?'), then I can claim I knew it. But if I require numerous prompts, perhaps I can't say I did. Moreover, some statements may require little prompting, not because you knew the answer, but because the answer is in some way obvious. The boy in Plato's Meno, then, cannot be said to have known what the length of the side of a square with twice the area of a given square was, as it took many leading questions to arrive at the answer. The answer to each question may be self-evident, but someone can't be said to know it unless they have tokened a mental state with that content before, and this has dispositional consequences. This discussion leads to an interesting empirical question: What role does non-conceptual content play in judgements of self-evident truths like spatial or mathematical ones? In other words, what is the role of experience in judgements that are generally taken to be *a priori* (Sloman, 2008)?

In relation to the discussion on the emergence of intentionality, there are two aspects of the account of how non-conceptual content contributes to conceptual contents that are relevant: it is diachronic, and results from work done (energy expended) by the subject. When it comes to concept possession, we can take a top-down approach (like Fodor), or bottom-up one (like the connectionists). Sometimes it may be right to take as our starting point the abstraction of language-like states (to treat these as atoms), e.g. when we want to explore the nature of systems that manipulate symbolic token with semantic interpretations, which we sometimes are, *ceteris paribus*. Fodor & Pylyshyn (1988) might have been right, to an extent, that the lower-level details are 'mere' implementation with respect to such systems. But, if we want to ask other questions, e.g. about how our concepts are formed, the implementation is important, because our concepts are formed from non-conceptual parts through time (on evolutionary, cultural and personal scales), these accumulations eventually forming generalizable concepts that allow us to group objects together usefully in kinds, which can be analysed in terms of topographical regions in conceptual space.

Much of the rest of what follows will be telling the story of how such concepts are formed, which will also function to explain why such mental kinds will not be reducible to statements that refer only to their physical realisations. Part of this story will address the question of what needs to be the case for us to say of something that there is something it is like to be that thing (Chapter 6: Physically Embodied Minds), and why such creatures are the kinds of things about which true intentional explanations are made.

Chapter 5: Embodied Agents

5.1 Rational agency

Before returning to whether reasons for an agent can be the kinds of causes referred to in scientific explanations, we should be clear about what acting for a reason amounts to. To make predictions about the behaviour of those we take to be agents, we assume rationality on their behalf (Heyes & Dickinson, 1990, p. 107; Dennett, 1987, p. 185), that is, we take them to be acting for reasons. This assumption is similar to those we make when predicting the behaviour of any object: we assume that there is something about the object that makes it like other things of the same kind, and which will cause it to behave like things of that kind do. In the case of rocks that amounts to, partly, assumptions about the behaviour of certain minerals in the conditions we find them. In the case of agents, it amounts to assuming agents will act in accordance with the reasons they have and their context, including normative standards: taking all your clothes off is irrational in sub-zero temperatures, even though it is explicable in the delusional stage of hypothermia. In this chapter, I will flesh out what these assumptions amount to by looking at what we mean by 'agent,' by 'action,' (which bears on which behaviours of agents we are interested in), and finally by 'rational' (which bears on what we are assuming about agents and their reasons for action).

An agent is a creature situated in an environment from which it receives information that is used as a basis for actions within that environment in order to achieve certain goals the system has (much more on this in §6.3 Panpsychism & Composition). In other people's words: 'An autonomous agent… is a system situated in, and part of, an environment, which senses that environment, and acts on it, over time, in pursuit of its own agenda' (Franklin, 2003, p. 51); or, 'Agents actively sense their environment, have aims that form the basis for decisions and plans, and act on them' (Aleksander & Dunmall, 2003, p. 8). What it is to sense the environment and to have that information available for use can be cashed out in terms of the tokening of representational states as detailed above (§4.2 Representations). The way that this sensing is shaped by the goals it is put to use in achieving, and in turn shapes those goals, will be further explored below (§5.2 Feedback and Feedforward).

If being a performer of actions is part of the definition of being an agent, we need an independent definition of action. Certain behaviours, like eyeball saccades that scan visual scenes to build representations of it, are not actions, as they are subpersonal and nonconceptual (Prinz, 2006, p. 14); they don't satisfy the generality constraint (Evans, 1982), and are not norm governed (McDowell, 1994). By contrast, behaviours that count as actions are driven by concept using mechanisms. In picking up a cup to drink from it, my behaviour is an action as it uses the representation of the cup and its contents, but the particular movements and processes involved in

moving my arm, aligning my hand, grasping, bringing the lip of the cup to my lip, etc., are not actions themselves. Whether there are some cases, e.g. animal actions, that can count as actions despite being non-conceptual we will return to below (§7.2.1 United We Stand).

Rationality is also defined in relation to actions, that is, a rational act is one done for a reason, and only actions performed for reasons can be called rational. The agent has mental states that contain information about the world, about the goal state of the world, and how to achieve that; what distinguishes such actions from triggered reactions is the normativity constraint: the agent's reasons can be wrong, for example having mistaken beliefs about how to bring about the goal state. Being based on faulty reasoning doesn't make an action irrational; the possibility of being wrong is necessary to being rational. For triggered reflexes, it isn't clear we can say they are wrong when they are triggered, since evolution may design them to be oversensitive if the costs of a certain proportion of 'false positives' are lower than those of a proportion of 'false negatives.'

Thus, behaviour that results from subpersonal, automatic causal processes is not rationally evaluable. Salivating caused by seeing and smelling food is not an action based on decisions of the agent's that can be said to be faulty. The drives behind actions, whether evolved or 'programmed,' act as 'motive generators' (Sloman, 1987), and may themselves be judged to be rational or not depending on whether they fit with other, broader goals of the system. So, given that a person has a shoe fetish, obtaining shoes is rational, in that it achieves their desire for shoes. But sexual desire related to shoes could be judged irrational in that it does not serve the broader aim of reproduction. Putting that aside, in deliberating, we imagine possible situations and actions and choose between them. Consciously deciding in this way is engaging in 'voluntary' action (Baars, 1997, p. 131). The success or failure of actions brought about as the result of conscious deliberations will inevitably inform future deliberations in relevantly similar situations.

For an action to be rational the beliefs behind it must be sensitive to the causal consequences of the action (Heyes & Dickinson, 1990, pp. 108-9). This means that simple observations of behaviour alone are insufficient warrant to attribute intentionality to creatures (Heyes & Dickinson, 1990, p. 112), and bars us from counting thermostats and other simple homeostatic mechanisms as intentional (Heyes & Dickinson, 1990, p. 113). There is a relativity here, though: the actions we perform become habit, and may not be immediately sensitive to changed environmental contingencies, but can still be said to be rational. We need to include a 'reasonableness' condition: a state is a belief if the creature would change it under conditions accepted as adequate to cause a change.

There is a difference between this position and Dennett's intentional stance: 'Whenever one can successfully adopt the intentional stance toward an object, I call that object an intentional system' (Dennett, 1981, p. 238). He says that in order to predict the behaviour of certain systems we assume they are rational, and being liable to be so described is all there is to being intentional. There is a vagueness here regarding how well and for how long a system has to behave in the way we would expect a rational system to behave, but let's say that we can, in the spirit of Turing, decide on a procedure, and let's call this passing the Dennett Test.

However, this still leaves us unable to distinguish between something's being rational and something's acting *as if* it were rational. In fact, it makes it meaningless to say that something is merely imitating rational behaviour if it does it well enough for long enough. In the case of the thermostat, Dennett assumes that taking the intentional stance towards it would not succeed in a sufficiently broad range of circumstances to count as a 'successful' adopting of the stance. But any organism, or robot, or cloud of interstellar gas that passes his test, just is rational. The same reasoning is used to dismiss the conceivability of philosophical zombies: anything that acts as if it were conscious, just is conscious.

Is it not conceivable that a simulation of conscious behaviour could be good enough to fool enough people enough of the time without actually being conscious? At what point does saying of a chess playing computer that it is trying to capture my queen become a literal rather than a metaphorical statement? There must be criteria for distinguishing the metaphorical from the metaphysical claim independent from the instrumental adoption of the intentional stance, with the instrumental being a guide to what we might apply those criteria to, but not a diktat about what to count as intentional. Whatever Turing-esque test we apply, there is always a chance that something will, by chance, pass it. The probability may be vanishingly small, so small that in real life it can be ignored, but in the world of thought experiments, this possibility is enough to say that we need other criteria to say why something is or is not truly rational. What such criteria may be will be looked at below (Chapter 6: Physically Embodied Minds); for now, we will continue to assume that it is having reasons that makes a rational agent, this 'having' being the tokening of the 'right kind' of mental state.

Hurley (2003) argues that it is in the realm of action rather than abstract, conceptualised, inferential thought that reasons are to be located. We can act for reasons that are not fully generalizable and not fully thought through, but are context bound, existing in little 'islands' of rationality rather than in the open 'space' of reasons (Hurley, 2003). Reasons are not, fundamentally, for justifying beliefs about what to do, but rather for guiding intentional action, and consequently, we should think in terms of conceptual *abilities* rather than conceptual *content* (Hurley, 2003, p. 232).

As mentioned, to be conceptual is to satisfy a generality constraint, which in terms of reasons for actions requires that the reasons satisfy normative constraints that apply generally, rather than to the immediate context alone (Hurley, 2003, p. 233). A creature has reasons it can call its own if, from its perspective and relative to other intentional states these could combine with due to the holism of such states, its action is reasonable given its context and the normative constraints in play. These normative constraints include some specification of the circumstances under which a creature should, if rational, alter its beliefs or desires.

There is a spectrum from pure stimulus-response, to context-free, 'inferentially promiscuous' rationality. Stimulus-response relations are invariable, and relatively insensitive to counterfactual changes in stimulus-response relations once conditioned, whereas intentional action is based on the holistic and normative relationship between means, ends, intentions and perceptions (Hurley, 2003). The fact that such flexibility is not all or nothing leads to the conclusion that creatures can act in ways that sufficiently satisfy the criteria of holism for intentional action, but who fall short of the fully flexible abilities required to be a true concept user. This leads one to ask how much flexibility is required to be said to have a concept, and how often, if ever, we achieve that level.

The Wason selection test (Wason, 1968) indicates that we don't naturally operate at the level of the fully general. This test shows that while we intuitively know how to check if a rule of social obligation is being violated, this ability does not generalise to checking violation of the same rule presented in a way devoid of social context. Take the social rule, 'If you help make the bread, you can help eat the bread.' We know that to check for violation of this rule we check that those who are eating it have helped to make it. But to check a rule like 'If a card has a picture on it then the other side will be blue,' which has no social connotations, many who have not studied formal logic will check any non-blue cards despite the fact that the appearance or not of a picture on the reverse side of such a card has no bearing on the rule. We may not, as Hurley (2003, p. 241) says, always exercise our conceptual abilities in a fully flexible way, but if we conclude from that that we should not really count ourselves as fully-fledged concept users, then the term would be robbed of its meaning. Therefore, we need a less strict definition of what it means to be rational.

As we noted, normativity requires the possibility of making mistakes, of not correctly applying a general rule that has been abstracted from previous instances. We have to be able to say that the agent mistakenly acted for a particular reason, rather than merely reinterpreting the reason to explain the action. In order to do this (i.e. to solve the rule-following problem of Kripke (1982), after Wittgenstein), we need criteria of correctness external to the organism that was acting (Hurley, 2003, p. 241). Complex organisms like us, which use representational feedforward and feedback

control mechanisms (see §5.2 Feedback and Feedforward), are embedded in an environment and have a teleological history (biological and social) that underpins functions and thus normativity (Hurley, 2003, p. 243). Reason and action take place in the world; we perceive the world and our actions in it, and the effects of these actions feed back into an updated perspective. In the case of fast, expert action, we create expectations to complete the loop internally. Such 'forward models' may also be used to predict what other modules may do, and what other agents may do, thus favouring a simulation model rather than a theory of mind to predict the behaviour of others (Hurley, 2003, p. 243). These kinds of simulation-driven processes are relatively context bound due to their being more a case of having skills rather than knowledge, know-how rather than know-that (Hurley, 2003, p. 251).

To be said to act *for* a reason and not just be describable as acting in agreement with reasons, that reason must be the cause of the action. Reasons don't have to be things that we are conscious of having, but they are still at the personal level, as they are normative and holistic, which provides a 'coarse recombinant structure' (Hurley, 2003, p. 234). Rather than existing in a 'space of reasons,' where all contents are laid out and the interconnections between them laid bare, there are 'islands of reasons emerging from a sea of causes…. language provides the bridges that finally link these islands together' (Hurley, 2003, p. 253).

In the coming sections we will add to this outline of mental kinds, filling in details of the story of how this process of emerging happens, and what conclusion this leads us to regarding the place of mental causation in a material world. A recurring theme will be that of feedback, this being a dynamic at the centre of many processes involved in the emergence of cognition.

5.2 Feedback and Feedforward

This section highlights the centrality of the circular causal dynamics of feedback. There are two ways to look at feedback in cognition: firstly, diachronically over biological, cultural and developmental timescales; secondly, synchronically, in lived experiencing of the world when acting in it, where the phenomenal character of our present experience is partly formed by the accumulated deposits of our experiences, individually, culturally, and biologically. This is in contrast with a naïve view of perception which takes the input as given and then applies knowledge to this input in processing. Rather, we see that the input is often formed by these contingencies, it comes *as* an experience of something, rather than being pure information that is discovered to be about something, post processing.

> …our capacity to relive or rekindle contentful events is the most important feature of consciousness…. this echoic capacity is due in large part to habits of self-stimulation that we pick up from human culture… the Joycean machine in our brains is a virtual machine made of memes. (Dennett, 2005, pp. 171-2)

An important feature of the kinds of feedback in play in cognition, I will argue, is that they lead to emergence of the kind that blocks reduction without violating physicalism. Feedback based on small differences, changes at bifurcation points, leads to the emergence of one thing rather than another. In the case of physical systems like fluids, small, random events, plus the background constraints of the surroundings, may lead to the emergence of phenomena like whirlpools. We can make general statements about the causal powers of whirlpools, and even though we may not be able to predict their emergence due to the indeterminism of the initial random event, we can explain those causal powers in terms of those of the individual particles caught up in that dynamic. However, in biological, cultural and cognitive systems, the explanation of why one form emerges rather than another is not purely physical, I will claim, due to the selective forces that act on them. The causal powers that evolved systems have are explicable in their own terms; once the mutation event causes a new kind to be born, the causal powers this kind of thing has 'slide over' the top of the sea of physical causes from which they have bubbled up. This allows us to distinguish metaphorical uses of intentional statements like 'The whirlpool is trying to pull me under', from literal uses in the case agents trying to drown you for some reason.

5.2.1 Evolution

The processes of evolution have built cognitive systems like us over a long period through the diachronic feedback that is natural selection. Understanding how this works, and the implications it has for our understanding, is key to the project of naturalising intentionality, by which I mean, telling a convincing story of how natural processes could result in beings like us, thus lancing the boil that is the air of mystery that generally surrounds conscious cognitive states. If each step on the journey from there to here is unmysterious, then we can conclude that there is no mystery regarding how we got here, even though, when looking back, we might wonder how we arrived in this place. This is nothing more than to say that we are on a continuum with other animals, which should lead us both to have more respect for animals, and to be less arrogant about our own position in the cognitive hierarchy (not that we are not at the top, just that we may not be so far above those below).

An important fact is that human kinds are subject to the triple interplay of biological, cultural & individual evolution. We inherit, among other things, genes from our ancestors, beliefs and artefacts from our cultural forebears, and memories from our earlier self. These processes create a

dynamic landscape of possibilities, analogous to three body dynamics in classical mechanics: simple parts creating complex and chaotic patterns. Joint attention is an example of the biological underpinnings of cultural evolution: we have evolved to be able to, and to want to, notice what others, particularly care-givers, are paying attention to, and to direct our attention to it (Heal, 2005). As a result, we learn important things about our surroundings, reliably, without needing to have that knowledge innately given. Furthermore, it may be that our attention is drawn to an aspect of the environment that we change in some way to suit us, and that with experience we might improve that manipulation and pass this on to those we come to care for. This is the hominid speciality of cumulative downstream niche construction (Clark, 1997). We inherit an environment that has been altered, and sometimes we inherit behaviours or artefacts that serve to construct an environment.

This leads to the question of what the correct units of selection are (Lewontin, 1970). Are they genes, traits, organisms, organisms together with their environments? Evolution works if something can produce copies of itself, which may vary to some degree, and which may lead to differential success in further copying. There is good reason, I believe, not to be reductionist about evolution, thinking that there is one unit of selection, namely chunks of DNA. Not only is it not so easy to identify bits of DNA with genes for traits, it is impossible to specify what genes are for independently of an environment: the same bits of DNA could lead to different traits in different environments, and the same traits could be realised by different strands of DNA (Hull, 1972). Moreover, there are many other kinds of things that satisfy the three conditions just given, a learnt behaviour being just one. All such processes involve a reproduction of something variable and selection pressures acting on them. These are feedback mechanisms, which set up attractors in the dynamic landscape, where the initial bifurcation may be based on a random factor. Anything that is a reliably recurring developmental resource (Mameli, 2004) can become part of this process, including an environment which is itself transformed through niche construction.

Selective processes are inherently hierarchical, in that the competition to produce more copies of something may happen at different levels, but that some levels are more basic in that they are necessary for the higher levels (Brandon, 2007). For example, an idea can be copied, can vary, and can lead to more or less copies of itself, but this can't happen without already existing means of storing, communicating and acting on ideas. If ideas are subject to evolutionary processes, and assuming ideas are realised in various ways in different organisms, then the best level of description for the causal process of evolution will be at the level of the thing being copied. If we described only the physical level processes that occur when teaching an AI an idea, the idea itself would be missing.

Sometimes the natural unit of selection may be an allele of DNA, the whole organism, a kind of behaviour, or even broader social structures, where there may be no independent measure of similarity among the units apart from that high level of description. Brandon (2007) goes on to say that the environment (which may be a social structure) cannot be considered as part of what is selected as it is not copied, but rather it is part of a constant background in which evolution happens. However, this distinction may not be as clear as it seems. Our language is part of our environment, but could be said to be subject to its own evolutionary process. The tools we use are reliably occurring developmental resources, but we copy and improve them. Sometimes, we inherit behaviours or artefacts that serve to construct an environment, and as this is then a stable part of the normal development of organisms like us, which can be copied with variations that lead to greater or lesser success, then that is subject to selection along with other traits of an organism. The distinction between environment and units of selection is interest relative in the same way as that between a cause the background conditions. As explained above (§3.1 Causation), the distinction between background conditions and causal processes depends on our explanatory interests.

Mameli (2004) tells a story to illustrate the idea that what gets copied may not just be bits of DNA: that of the lucky butterfly. Imagine a species of butterfly that lays its eggs on a particular kind of leaf so that its offspring have a good food source when they emerge. This is achieved due to a mechanism that imprints on the kind of leaf a young butterfly first encounters, allowing it to re-identify the right kind of leaf when it comes time to lay its own batch. Sometimes, random errors enter into such processes, and occasionally eggs are deposited on the wrong kind of leaf. Most of the time this would probably be a disaster for the hatchlings, leading them to not flourish and so not to have any descendants. However, if by chance the eggs land on a leaf that is more nutritious than the original kind, the descendants of this butterfly will flourish, and they will now be imprinting on the new kind of leaf. This could be called a speciation event, as the descendants of this lineage may be quite different merely due to the difference in nutrition, pigmentation, etc. that they receive from this new kind of leaf, and this in turn may lead to significant differences in behaviour. In this case, the environment is not just a given background; the genes of the new species have not altered, the imprinting mechanism is the same: what has been selected is the kind of leaf.

Similar arguments about externalising the units of selection apply in the case of cultural evolution. Culture contains developmental resources that get passed on, like the knowledge of what to eat and how to get it, which are a form of extragenetic inheritance. As cultural creatures, we have the mechanisms (e.g. mindreading, expectation) that allow us to take advantage of this, and which,

because of their lack of content, permit the evolution of new behaviours and construction of new niches (Mameli, 2001, pp. 601-9).

These abilities have an evolutionary advantage over genetic selection. In social animals that have evolved the ability to learn behaviours from their conspecifics, advantageous behaviours can rapidly spread through a population without needing to wait for random genetic mutation to hit upon the solution. Moreover, given that a new, advantageous behavioural trait that an organism develops in interaction with its environment becomes entrenched, those that can learn it more quickly and accurately will be more successful, and in time genes and/or cultural resources that help in this process may be selected for. This is the Baldwin effect (Baldwin, 1896), where selective pressures work on a general learning ability so that possible behaviours can be tried out and passed on. This will lead to the need to have a means of communicating the content of one's domain general cognition, e.g. language. If this is the case, then Lamarck was not so wrong: although giraffes cannot pass on longer necks to their offspring from a lifetime of stretching, humans can pass on a method of climbing smooth trees to get at the honey, and an ability to learn it.

The possibility of non-genetic inheritance, and the multiplicity of selective mechanisms this opens up, means that evolution can act on many levels, and also that there can exist a variety of property types that have an ontological status comparable to that of the properties of 'traditionally' evolved biological organisms. Moreover, the 'essence' of many of these properties may not reside in purely individual, internal facts, but in facts about the social environment. One internal fact must be true though, and that is that we have evolved to be the kinds of creatures that are fundamentally flexible, so that our culture can determine how we develop. This is one reason for having a long maturation process: to allow time for the process of individual development to shape us.

We are plastic people. Our success as a species depends on our ability to become functioning adults no matter which of the earth's environments we are born in, and to be able to rapidly spread innovations horizontally through and among populations, which then improve on them, and so on (this is Tomasello's ratchet (Tomasello, 1999)). This view of humans is another refinement in the rationalism vs. empiricism debate. I take it as rebalancing the argument in support of the empiricist, against rationalist cognitivism's rejection of behaviourism. Not that we are born as 'blank sheets,' the genetic inheritance is necessary, as are environmental factors; rather than a painter with a blank sheet, we are like an environmental artist working with the materials at hand, constrained by our tools and materials, but free to create new and interesting objects from those. A good reason to be excited about this way of viewing human's place in nature, is how it allows analytical, scientific philosophers to take seriously more interpretative, social treatments of human nature, like

Vygotsky's idea of the importance of internalising cognitive structures in the process of learning within a culture (Vygotsky, 1978), or Bourdieu's notion of 'habitus' (Bourdieu, 1980), this being a social organisation which determines what positions we may take and shapes our self-concept. Once again, feedback plays a central role in how we develop our behaviour as individuals, both diachronically through positive and negative feedback, and synchronically through the way we experience the world.

Organisms that can respond to their environment flexibly can succeed in environments that change. Sense organs that respond to a narrow range of stimuli in a fixed and fast manner linked directly to particular actions are obviously very useful. But, if an organism has multiple sensory channels, and needs to coordinate these inputs in a way that is flexible rather than reflex (and assuming it has the luxury of time to think things through), it needs an ability to put these inputs together in a less domain specific way and decide which action to take. There must be a space for general reasoning which is not linked automatically to any particular action, and which can represent aspects of the environment 'offline,' without the presence of stimulus to the senses. This requires the ability to have in one's head representations for things in the world, and the ability to generalise, i.e. the capacity to have concepts. Dennett's 'Popperian' creatures are such that they are able to use information to form models that allow them to think through possible actions before committing to them, thus giving them an adaptive advantage over creatures that have to rely on the genetic feedback of trying actions and seeing who survives (Dennett, 1996; Clark & Grush, 1999). Moreover, this information may come in the form of linguistic communication from others.

Carruthers (2002) argues that it is unlikely that language is necessary for complex, domain general thought (problem solving abilities that can be used in various contexts), since pre-linguistic infants and non-linguistic animals exhibit such abilities, although to a limited extent. Furthermore, he says that it is the domain general thinking skills that evolved first, language being a later refinement that required the pre-adaptation of the domain general cognitive abilities. But he also says that being a language user is necessary for cross-modular thinking, which is where the processes of peripheral modules not normally accessible to other processes can be brought together.

There seems to be a ratchet effect between domain general thought and language use that makes it difficult to separate them and say which one is necessary for the other. Some kind of domain general conceptual abilities may be necessary for language to start, but learning a language itself moulds our minds into general reasoning machines in a way that wouldn't happen without exposure to it during development, and as these abilities become more important in our evolutionary niche, there is selective pressure to be better at picking up language and domain general cognitive skills.

What genetic resources we have at the present stage of our evolution that grounds these learning abilities is still an open empirical question.

To explain the causal properties of biological and cognitive kinds, we need to refer to the process of variation, reproduction and selection that produced them. These processes may be happening on multiple, interacting levels, e.g. biological, social, and developmental. These are feedback mechanisms, which set up attractors, where the initial bifurcation is based on a random factor and an environment which is itself transformed through niche construction. Anything that is reliably recurring and plays a role in development (not just genes) can become part of this process (Mameli, 2001, pp. 614-617).

Therefore, despite the fact that biological and cognitive systems are wholly composed of physical stuff, descriptions on the purely physical level will miss the patterns describable at the higher level. Evolved organisms have the causal powers they do because of the accumulation of selection events, and things that owe their design to the same selective forces will have similar causal powers, allowing us to make general statements about things of that kind. These causal powers are emergent properties of the whole situated organism, constrained by the physical stuff they are made of and surrounded by, but not definable by reference to the physical stuff they are made of alone, due to the multiple-realizability of those causal properties, and the fact that the same, narrowly described physical embodiment might have different causal properties in other contexts. This is not to say that there is anything non-physical happening, just that the only way to draw the line around those parts of the world that are relevant to a particular explanation is to do so using categories that are not part of physical science; the objects referred to will not be defined by shared physical essences. Evolutionary explanations (Darwinian or otherwise) are like this. An explanation of cake eating behaviour in someone with no need for more sugar in their diet, for example, will refer to mechanisms that evolved in a low-sugar environment. This explanation may be supplemented by other levels of understanding regarding the physical realisation of the mechanism and how it functions, but those lower-level, micro-explanations will not be 'bound together' without the higher-level explanation.

5.2.2 Expectations

Of the kinds of feedback more immediately involved in our phenomenal mental lives, expectations (first person predictions) have received a lot of attention recently (e.g. (Chrisley, 2009; Clowes & Chrisley, 2012; Clark, 2016)). In standard models of cognition, the line of causation from perceptual input to experience and ultimately to action is one-way, where the mind is active in performing inferences on the information in the input. In systems that actively use predictions, the perceptual

input is shaped by information from within us in the form of expectations, therefore causation is partly circular. An implication of this is that the conceptual maps we create to refer to when navigating the world may result in changes being brought about in the world due to our actions.

In this way, expectations can be self-fulfilling (Mameli, 2001, p. 609). This makes it impossible to cleanly disentangle the causal lines between the world and our understanding, which may lead some to the anti-realist conclusion that the concepts used to refer to significant features on our maps are a product of the mind rather than the world (Gärdenfors, 2000). I disagree, for, none of this means that there can't be a determinate answer to the question of how the world needs to be in order for our judgements about it to be true, even if there were other possibilities for how our judgements would be, and how the world would be given that our expectations could affect it over time. Before giving details to support this assertion, I will clarify the concept of expectation in use.

> Expectation is, grammatically, a state; like: being of an opinion, hoping for something, knowing something, being able to do something. But in order to understand the grammar of these states it is necessary to ask: 'What counts as a criterion for anyone's being in such a state?' (States of hardness, of weight, of fitting.) (Wittgenstein, 1953, p. 572)

We are not necessarily referring to the state of being expectant, like when you are waiting for something. Of such states, we can ask, 'What makes the statement "The dog is expecting a bone" true?' We might answer by saying that for it to be true, the dog requires a representation of the bone which can be used to imagine a future situation that contains a bone, and that it would be able to recognise as a bone a variety of possible instantiations of the type. In this case, it could be said to, at least, possess the recognitional concept of bone.

I am more concerned, presently, with the way our moment-by-moment experience is partly determined by expectancy effects. The idea that our understanding of the world, possibly in non-conceptual form based on either experience or innate resources, may affect how we experience the world to be, is not new, but it has recently had a lot of attention paid to it. Gibson (1979) introduced affordances to modern cognitive science, but this idea had long been of interest to phenomenologists (e.g. Heidegger (1927)). Ryle (1946) was talking about such skilful knowledge of the world when he distinguished 'knowing-how' from 'knowing-that.' More recently in cognitive science is the enactive approach, which sees our perceptions of the world as including the possibilities for action, or 'sensory motor contingencies' (O'Regan & Noë, 2001). This makes evolutionary sense, since this allows us, through learning a skill, to set up automatic routines that are quick and efficient (until the world turns out not to live up to expectations); we become 'coupled' to objects in a way that negates the need to go through chains of inference from perception to action, or indeed the

need to internally store information about the behaviour of kinds of objects or aspects of the world (O'Regan, 1992). Knowledge of such aspects of the world may not be stored in terms of propositions, but they are internally stored in such a way that this knowledge is available to guide our actions.

However, I don't want to endorse a radically enactive position that sees everything in terms of 'action potentials.' Such 'enactive machines' that are either internalised through development, or hardwired through evolution, are but a part of mental architecture that determines our experience of the world and our actions in it. For one, if all action were performed in such tight couplings with perception there would be no room for deliberation; everything would be reactive, driven by input. Of course, over time there is room for negative and positive feedback to influence the forming of future expectations, but deliberate action is a more immediate break in the loop. Furthermore, enaction seems to require that parts of the world are causally involved in our interactions with them, but this requires them to be causally proximal. If we are considering thoughts about things that are remote, or even not present, we still need to rely on internally stored, representational resources. Similarly, if we are removed to an environment radically unfamiliar to us, we need to have a way of experiencing and acting on the world not guided by expectations alone. (Having no correct expectations about events in your surroundings may be what accounts for the state of being in culture shock.)

Having said that, an important and relevant upshot of the kinds of feedback and feedforward systems involved in 'normal' action in the world is that they may provide an explanation for the existence of conscious awareness of mental states. Feedback is where information which tells the system its current state informs decisions about what actions to take to try to reach a desired state, which is then fed back into the system and checked again, in a loop. Feedforward is quicker, merely taking the current state (including environmental factors) and, projecting a future, desired state, and calculating what needs to be done to reach that state. Visual feedback is too slow for rapid body movements, so a feedforward model, based on learnt contingencies, must be used. 'Efference' copies of the action are sent to other brain systems for various uses. Both require 'models' of the system and any aspects of the environment that affect control; Holland & Goodman (2003, p. 80) call these 'adaptive model-based predictive controllers.'

Consciousness may have the evolutionary role of checking actuality against expectations, the results being fed back into future expectations: 'the internal feedback made possible by the efference copy routes is a vital feature of the neural processes underlying consciousness' (Cotterill, 2003, p. 32).

Also, efferent copies may trigger memories of previous outcomes of such actions without the actions being carried out (imagination). Interestingly, the neural routes involved in this appear to be the same as those involved in directing attention. Short-term memory of where the creature is within the virtual landscape of possible actions is necessary for navigating the options afforded, patterns emerging in these closed-loop circuits as chaotic attractors, self-organising patterns that are structurally stable (Cotterill, 2003, p. 42). So, consciousness may be this ability of a creature to model itself and its relation to the world, where there is a clear difference between representations of the inner and the outer (Holland & Goodman, 2003, p. 86). Its evolutionary advantage is to be able to vary behaviour so as to bring about desired outcomes. Models can be used by a creature to distinguish what it could, should and would do, that is, what is possible, what is best, or what may need to be deferred (Holland & Goodman, 2003, pp. 98-101). This critical self-awareness is achieved through a form of re-entrant feedback where we can focus attention on the contents of our perceptual systems and other cognitive processes (Stuss, 1991). There are multiple such mechanisms for focussing on what our brains are doing, and learning from this. For example, vision receives input from higher areas, such as knowledge representation (Edelman, 1989), which can lead to a phenomenon well-known in the philosophy of science, namely the theory-ladenness of observation (Kuhn, 1962). We build our sense of self though this critical process of inner attention, this selfhood then becoming part of the critical process of development (Steels, 2003, p. 183).

Also relevant to the current account are the implications of this for the nature of mental causation, in that it includes, essentially, a historical element in the form of past experiences and thoughts. Of course, this experience is somehow realised in the physical brain, but referring merely to the physical instantiation will not enable us to answer all the interesting 'why' questions about mental states. Firstly, as mentioned, we are talking about structurally stable, emergent patterns, and small physical changes do not generally make a difference, thus the micro-causes relevant to the realising parts are not normally relevant to the causal properties of the whole. Agents like those we are discussing are multiply realisable, which means we should look to the level of what is being implemented in order to find causal generalisations. What is being implemented is a way of making the past bear on the present, and to understand why a particular action was carried out, we need to understand the way in which past experiences form present inner reality.

5.3 Externalism

We have seen hints in previous sections that the supervenience base of mental states may not just include the brain. This is the externalist hypothesis, another word beginning with 'e' that says

something about the constitutive relation between brain, body and environment, situating cognition in space and time rather than viewing it as an abstract computational process the implementation of which is unimportant. Above, I have been arguing that cognitive states supervene on spatially and temporarily extended portions of the world, and so the physical embodiment, as well as the social embeddedness, of these states is important to understanding their causal properties. My aim has been to show how this broadening of the supervenience base of mental states blocks the kind of argument Kim uses against emergentism.

Enactivism claims that experience emerges from tight, skilled coupling between a subject and an environment, rather than being a matter of having abstract internal representations. Noë (2006) claims that perceptual contents, e.g. detail, three-dimensionality, colour, are present in experience not as represented, but rather as available. In this sense, experience has the content it does at a moment in time only as a potentiality: perceptual experience is a temporally extended activity of skilful probing of the world. The world is available to our reach, given our skill, and experience, being comprised of aspects of mind and world. Experience, it is claimed, isn't something that happens in us, it is something we do.

Noë (2004) uses the concept of the 'virtual' to clarify the way in which the world is present in our experience of it, by which he means the world is available given the 'sensorimotor contingencies' of the subject. So, something that is not in view, like the backside of a tomato, is virtually present in our experience of the tomato because we know that if we go to pick it up it will be roughly spherical. Moreover, it is not the case that you could separate the virtual content from the directly given content: 'Experiential presence is virtual all the way in' (Noë, 2004, p. 216). The world is present in experience virtually thanks to our online, dynamic access to it, and qualities are available in experience as possibilities and potentialities, but not as givens. Experience is a dynamic process of navigating the pathways of these possibilities, and as such our current visual representation of the world is not a function of just the information being sent to the brain down the optic nerves: my phenomenal experience expands my immediate horizons and takes me beyond myself to the world. This is, according to Noë, a pervasive feature of our perceptual lives.

A couple of upshots of this view are immediately apparent. As cognitive systems emerge through tight coupling between the brain and the world, and these dynamic systems are not to be understood as the rule based manipulations of mental symbols between inputs and outputs, they are relational rather than internal and abstract, thus avoiding the 'symbol grounding problem.' This is the problem faced by computationalism, raised by Searle (1980), that, if cognition is the manipulation of symbols according to syntactic rules, then they do not carry with them any real

meaning for a cognisor to understand, and if understanding is supplied by an interpreter of symbols, an infinite regress threatens (this is another kind of homuncular objection). Another upshot, and one we will return to (§6.2.1 The Feeling of Things), is that not being abstract and self-contained, but rather constituted by patterns of bodily action in the world, the form of embodiment will determine the type and character of cognition. Before exploring these claims, I will clarify some terms used in the debate.

5.3.1 Physical Bodies in a Social World

When it comes to giving reasons for action, externalism requires not just internal facts but also 'exogenous constraints' to be present (Hurley, 2003, p. 243). The words embodied and embedded (or situated) denote cognitive theories that see mental states as arising from the physiology of the brain and body, rather than the body just providing the channels through which information enters the computational system that is realised by the brain. Moreover, embodiment theorists see the interaction of brain, body and the world as being essential components in the emergence of intelligent behaviour, rather than external facts being the background against which actions happen (Clark, 1997; Brooks, 1991; Lakoff & Johnson, 1999; Varela, et al., 1991).

As an illustration: the cognitive processes involved in riding a bike are embodied due to the tight coupling between brain, body and bike. A classical computational cognitivist approach would posit something like an internal model of the body on a bike which is used to compute the next action given the information being received about the bike through the body. Rather than this input-process-output picture, embodied cognition sees the cognitive process as involving brain-body-bike: after learning to ride a bike (which requires getting on a physical bike; being told or shown won't work), you gain a skill in which the bike becomes an extension of you, and is a proper part of the description of the cognitive processes involved: you form a dynamic unity with the machine. In other words, a detailed representation of the bike is not a part of the processes involved; the bike is there to play the role of storing information about itself used in the process of being ridden. Information about, for example, the position of the gears is available in the bike, not represented internally.

Cognition may also extend into the world in terms of being culturally embedded (we saw an illustration of this in §4.1 Mental Kinds). Sometimes actions can only be fully explained by reference to the social structures that give them meaning. These will have been internalised as model-like structures to some extent, but understanding them will require referring to the process of enculturation, and other causal factors that exist outside of the subject, within his or her cultural

world. Such actions rely partly on the internal motivational factors common to humanity, like fear of being shamed, that are driven by the importance of maintaining a certain position within a group, but the particular form they take may require active motivating, triggering and moulding by other actors in the group, or even symbolic prompts.

Wilson (2002) has a useful list of properties taken to hold of cognition according to the embodied approach:

1) Cognition is situated: it happens in the world and involves perception and action.
2) Cognition is time-pressured: it needs to be understood in terms of what can be done with the available resources in the time given. It uses 'cheap tricks' to achieve ends, such as rough-and-ready rules of thumb, rather than costly computations.
3) Cognitive work is off-loaded onto the environment: we design the environment to cue us, or store information in it for use when needed.
4) The environment is part of the cognitive system. Parts of the environment participate in the flow of information in such a way that they are best seen as part of the cognitive system, rather than merely as a source of inputs, so if we want to understand the organisation and function of cognitive systems, we need to include the environment. A system is something more than a mere aggregate; the properties of the parts must be affected by their participation in the whole. Cognitive systems have properties of both facultative systems (temporary arrangements of matter with vague, interest relative boundaries (e.g. ecological systems)) and obligate systems (more permanent arrangements relative to the lifetime of their parts) particularly as we move from one context to another.
5) Cognition is for action: the subsystems (e.g. vision and memory) must be understood in terms of how they contribute to situation specific behaviour.
6) Off-line cognition is body based: it is based in mechanisms whose function is to guide action in the world, e.g. sensory processing and motor control.

There are clarifications required for some of these points, particularly the distinction between facultative and obligate as this seems relative to the interests of the observer, too. We will return to this in the following, but for now, let me reiterate the purpose of putting forward embodied/embedded cognition in the context of this text: if it is the case that 'empirical externalism' (to contrast it with content externalism) is true, that is, if, in order to make sound generalisations about the actions of human agents it is necessary to include facts about processes

and events not spatially or temporally local to the brain of the subject, then the arguments put forward by Kim in favour of physicalist reductionism do not go through.

Later (§6.2.1 The Feeling of Things), we will see that viewing cognition as embodied also explains important features of phenomenology, but the major motivation behind embodied theories is economic. Since cognition 'in the wild' is time-pressured, and geared toward action, cognitive work is off-loaded onto the environment. In itself, this may not be something that should necessarily worry someone who sees cognition as a fully brain-based activity, as it could just be seen as part of developmental psychology, leading to an interesting reconfiguration of what we think goes on in the brain (e.g. the effects of expectation and attention on what is represented), rather than a radical new way of looking at how cognition itself happens. For a more wide-ranging change to our view of cognition, the environment should be seen as part of the cognitive system in a constitutive way, rather than just as something that is occasionally used in cognitive processes (not just as a tool, but as a proper part).

What does it take for something to be a proper part of a cognitive process? Inputs are not traditionally seen as parts of cognitive processes (remember methodological solipsism). These words are inputting concepts into your cognitive system: you extract the information represented in them using your knowledge of English; this causes tokenings of concepts which are then available for cognitive use. When concepts are tokened due to internal processes, like a chain of thought, this is different: there is no equivalent of reading to decode the information contained.

This distinction between being inside and outside the mind makes sense to the extent that it allows us to see ourselves as spatio-temporally located beings without 'over-extending' ourselves into the world. But, we should not be too ready to equate the proper parts of cognition with the brain itself. After all, the brain is just a certain arrangement of atoms like the rest of the world. It may be that some parts of the brain are only inputs to cognition proper, and that some parts of the world are not merely sources of inputs, but parts. Think of the visual system: it can go without detailed internal representations since the environment can serve as a repository of information about itself if it is immediately accessible. It is possible that our experience depends not only on what is represented in our brains, but on dynamic interaction between brain, body and environment.

Before continuing we should distinguish two ways the mind can be seen as being 'stretched' into the world. Vehicle externalism is the idea that objects in the environment that are utilized in cognitive processes in a habitual and reliable way count as a proper part of cognition rather than just an input to it (Clark & Chalmers, 1998). Content externalism is the position that the contents of at least some

of one's mental states are dependent in part on their relationship to the external world (Burge, 1986). Here we are mainly concerned with issues regarding causal properties of mental states and how they are physically realised, rather than issues of pure content alone.

Clark and Chalmers (1998) investigate our intuitions about what counts as inside the mind through the case of Otto and his notebook. Otto cannot form new long-term memories (anterograde amnesia); instead he uses a notebook to keep information he thinks might be useful. So, when someone asks him if he knows the way to the museum, he can get out his notebook, find the information, and answer, 'Yes.' The question being asked through this thought experiment is whether there is a principled difference between Otto's case, and that of a taxi driver with the Knowledge. Clark claims that if the action of reaching for the notebook is automatic and the information readily accessible (Clark & Chalmers, 1998, p. 17), then the notebook works like memory and so should be counted as a proper part of the cognitive process of recalling something one knows.

However, there seem to be good reasons for distinguishing the kind of access to the information in each case. Getting the information from the notebook requires finding the book, opening it at the right page, recognising the symbolic description and decoding it using learnt linguistic abilities, etc. In the case of mental memory, the information should just be there, 'at hand,' when needed. Admittedly, information stored in memory does not always reveal itself so readily; sometimes mental effort is required to access it. Moreover, remembering might not be an innate skill, as we can learn techniques that allow us to improve our ability to retrieve information. The distinction nevertheless seems real and meaningful.

The difference is not just that one process happens inside the skull and the other doesn't. We can imagine replacing a part of our normal brain with a 'box of tricks' located outside the body but hooked up to the brain 'in the right way,' so that the phenomenology of using that box is indistinguishable from the normal processes. Conversely, we can imagine an organ located inside the brain that would nevertheless count as outside the mind, like toes are. The difference, then, is in the immediacy of the experience of retrieving the information: normal remembering doesn't involve a process in which the information is coded and decoded, as in when it is written on and later read off a page. Such coding/decoding processes are necessary when the 'bandwidth' of the channel through which the information passes is not sufficiently wide. When systems are connected via sufficiently high bandwidth information channels, then 'new systematic wholes' (Clark, 2008, pp. 32-3) can be formed.

Below (§6.3.4 In Two Minds), I will expand on these thoughts while looking at the case of split-brain patients, who have had the connective tissue between the two cerebral hemispheres (the corpus collossum) cut. The role of the corpus collossum is to be a sufficiently high-bandwidth communication channel between the hemispheres, and here we have evidence for the fact that entities that could be regarded as separate subjects (the individual hemispheres) can become composite subjects, and we have an explanation of how this can happen: through their being connected by sufficiently high bandwidth information channels into 'systemic wholes.' In order to judge that the two hemispheres come together as one distributed subject, and indeed that the separated hemispheres can be subjects in their own right, requires an account of what it is to be a subject, which we will also return to below (§6.3.4 In Two Minds).

Husserl said we can understand others as subjects because we experience their body as a living thing (*Leib*) rather than a mere object (*Körper*) (Zahavi, 1994). Merleau-Ponty (1945) also stated that we don't infer the intentions of others from observations of their behaviour, rather, we find those intentions in ourselves when prompted by the observation of the actions of others. Gallese's (2001) work suggests that we are designed to respond to the goal-directed actions (and not just particular physical behaviour) of others specifically, and to mirror that behaviour, through the use of simulation (efference copies used in forward models), utilising so-called 'mirror-neurons' (Gallese, 2001, p. 43). Simulations are used in understanding others by 'putting ourselves in their shoes' (Gallese, 2001, p. 42). We don't need to construct theories (as in the 'child-as-scientist' view, (Gopnik & Meltzoff, 1997) using propositional attitudes in computational processes, or to have this 'theory of mind' in an innate module (Baron-Cohen, 1995). In contrast, accounts that rely on embodied empathising rather than theorising is 'direct and automatic,' not inferential (Gallese, 2001, pp. 42-44).

Gopnik and Meltzoff (1997) talk about an innate body schema that enables us to interpret the actions of others by mapping them to one's own motor systems. In support of this, evidence is given by Meltzoff and Moore (1977, 1994, cited in Gallagher (2001, p. 87)) that infants can sense what things in the world are similar to themselves, and map the behaviour of those things onto their own body schema in order to imitate them. This is not simulation, as we do not represent the body of the other to ourselves and infer intentions from that model, we experience others intentions directly through a shared body schema that maps perceptions to motor behaviour, without the need for an intermediate mental state. Rather, the body schema is 'a set of pragmatic (action oriented) capabilities embodied in the developing nervous system' (Gallagher, 2001, p. 87).

Gallagher (2001) claims that the kind of direct, primary intersubjectivity he is advocating is not merely the developmental starting point out of which mature (inferential or simulation driven) mind-reading abilities develop, but is the primary way we continue to understand the intentions of others in our everyday lives. Furthermore, he argues that not only in understanding the actions of others, but also in acting ourselves, we may be driven by the practical, habitual knowledge that 'the situation is just such that this is the action that is called for,' rather than a 'well-formed' mental state (Gallagher, 2001, p. 95).

By 'well-formed' mental state I take him to mean an explicit belief. However, even if we grant him this, we cannot conclude that actions are not driven by such mental states, or that we don't have the more abstract ability to understand others in terms of their mental states, through forms of indirect inference, given enough time for reflection and the tools with which to understand them. In fact, this higher-level ability may be necessary as a way of communicating more complex mind-reading abilities, and as such could form the foundation of scientific attributions of mental states that explain actions by classifying those states according to generalisations in law-like statements. In fact, it may be that this level of understanding, less direct, more reflective, is necessary to counteract the negatively habitual interpretations of others that we are prone to pick up as members of a society, thus itself playing its role in the cultural evolution of learnt mind-reading abilities.

Those pernicious habitual social attitudes, like prejudice, that flourish when passed on unreflectively, though, are rather like our appetites for salt: the consequence of an adaptation that was advantageous in the environment in which it evolved, but which now has some negative consequences. Cognition is 'scaffolded' by the socio-cultural environment, which follows its own evolutionary dynamic and which we internalise during maturation, and this is an efficient way of creating niches that nurture the cognitive capacities necessary to flourishing. As social creatures, we form a dynamic system where we observe and learn from others, forming a self that in its turn affects others (c.f. Mead (1934)). In the hunter-gatherer past, where positions in the group were clear and groups were in competition for limited resources, the 'othering' of people outside one's own group was probably adaptive: these attitudes sustained the kin-group, and the kin-group sustained the individuals, and the kin gene-line was maintained.

In the complex society that now forms an essential aspect of our developmental environment, we need an understanding of the causal influence this has on our cognitive processes, which requires a metaphysics of causation that has a place for upward, downward and historical causation, to allow

for the causal influences that the social roles we occupy exert on our action through the inculcation of dispositions to behave in certain ways:

> …the relational or structural approach that [the concept of the field] introduces is associated with a dispositionalist philosophy, which breaks with the finalism, allied to a naïve intentionalism, which sees agents as rational calculators seeking not so much the truth as the social profits accruing to those who appear to have discovered it. (Bourdieu, 2004, p. 33)

Society defines the roles we can take in it, and we internalise the possibilities for action afforded by being in that role: 'A scientist is a scientific field made flesh, an agent whose cognitive structures are homologous with the structure of the field and, as a consequence, constantly adjusted to the expectations inscribed in the field' (Bourdieu, 2004, p. 41). This structure is reproduced through us, utilising the ability we have to learn concepts, which bring with them generalising powers and (self-fulfilling) expectations: '…power relations are set up and exerted in particular through cognitive and communicative relations… [and] can be exerted only on agents who possess the categories of perception necessary to know it and recognise it' (Bourdieu, 2004, p. 55).

The investigation of this social-evolutionary development, and the mental structures through which it is maintained and exerts its influence, is beyond the scope of the current text, but I think this is another exciting space for a rapprochement between the structuralist sociology of the continental tradition and the analytical tradition that led to cognitivism. Having argued that the environment that provides the conditions for our development as individual cognisors includes the social as well as the physical, and that the first-person experiencing subject is central to the question of what is properly a part of us as cognitive beings (given that our understanding of others, which feeds back into our understanding of ourselves and therefore how we act, is not knowledge that is extracted through abstract inference), I will return now to the issue of how to determine which portions of the physical world are to be counted as constitutive of cognition, and which are background conditions that may have a causal influence on cognition, but do not count as part of the embodiment of the individual experiencer. For the answer to this question clearly has a bearing on the matter of mental causation.

Honderich (2006) has a radically externalist claim about the relationship between experience and the physical world. He puts forward some interesting arguments, using assumptions some of which I endorse, and others which, in my opinion, are shown to be problematic through his following them to his conclusions, these assumptions being ones used, sometimes implicitly, by the main targets of this work, namely intentional reductionists. Firstly, he claims that neurons (or by implication other physical substrates) are not part of my experience, because my experience is fully 'transparent' to me and I cannot be mistaken about what is and is not part of my experience: 'with respect to

consciousness, there is no difference between appearance and reality' (Honderich, 2006, p. 5). Next, he denies that intentionality is the mark of the mental: 'There wasn't a relationship of intentionality, aboutness or directedness in your consciousness of the page' (Honderich, 2006, p. 5). For him, being perceptually conscious of something is not a matter of there being a mental state that represents that thing, but rather for 'an extra-cranial state of affairs to exist — for there to be a spatio-temporal set of things with a dependence on another extra-cranial state of affairs and also on what is in a particular cranium' (Honderich, 2006, p. 6). Both parts of the world, inside and outside the skull, are needed for there to be consciousness of something: 'There is not much of a liberty taken in speaking of there being pages in both a world of perceptual consciousness and in the perceived physical world, and indeed in referring to each of a related pair of things as a page' (Honderich, 2006, p. 7). Honderich says his radical externalism denies the 'worn story' that the brain contains causally sufficient conditions for perceptual consciousness, instead claiming that neural facts are merely necessary (Honderich, 2006, p. 8).

Much of this is unobjectionable: for there to be perceptual experience of a page, there has to be a page and an experience; the same neural events without the page would be a hallucinatory experience of a page (I am ignoring for simplicity the claim that hallucination is a kind of perception). But it is problematic to speak of there being a page formed by the relationship between the pages in consciousness and in the physical world. It is more natural to assume there is the perceptual experience of a page, and there is a perceived page in the world, and nothing more. What difference to the perceptual experience of a page would the removal of the physical page from the picture make? It would make it a non-veridical perception, but that is a matter of judgement from outside the experience.

The false premise in Honderich's argument is, in my opinion, the assumption of the transparency of our experience. It may be that two experiences that appear to be the same to the perceiver, could in fact be different experiences. That doesn't mean that in the two cases there aren't aspects of the experiences that in principle one's attention could be drawn to that would show them to be distinct; but if this is done, then the experience is not the same anymore, so this does not show that the two experiences were in fact always different. There may be parts of my experience that I am not aware of, and I may be wrong about what I think I am experiencing. In other words, how my experience appears to me might not be how my experience really is. To argue *a priori* that this is self-contradictory is to beg the question, and only seems to be a reasonable reply because of the Cartesian heritage. But Descartes thought that the idea of an unconscious idea was self-contradictory.

Manzotti (2006) uses rainbows to show that consciousness extends into the world using a different argument, in some ways opposite to the above: rather than the objects of experience being partly constitutive of the actual experience, it is consciousness that is partly constitutive of external objects. Rainbows require particular conditions to occur in the atmosphere and in the observer. The rainbow example is supposed to show that the existence of the thing that causes the experience (the collection of reflecting drops taken as a whole) is abstract and un-unified until the effect has actually occurred in the observer. Therefore, causal properties depend on the causal network as a whole, rather than being located in a particular external and independent object; causes and effects become different ways of looking at processes.

As with the previous argument, much of this is unobjectionable, particularly the idea that picking out causes and effects depends in part on the point of view an observer takes towards a process, but I think he puts too much weight on the example of the rainbow. Rainbows depend on the point of view of the observer in a way that objects proper don't: rainbows are optical phenomena rather than things. Whenever sunlight hits air filled with water droplets, light gets scattered, reflected and refracted. An observer will see the spectrum painted in an arc because of the way this light and her position interact, even though light is going in all directions. Another observer stood next to the first will also see a rainbow, but there seems to be little reason to say that he is seeing the same rainbow as her, as his rainbow is in a slightly different place. This becomes clearer the further apart the two observers are. It could be argued that the same is true for any distally perceived object; after all, we are just seeing the reflection of light from molecules loosely bound together. However, the parts that make up the rainbow are much less bound to each other than those of a distant object, say a wildebeest. The parts of the wildebeest have a history, and a future, irrespective of being observed. It munches its grass and is chased by lions whether or not we watch; it is subject to processes that define its identity in a way that rainbows are not, that is, independently of an observer.

Although the above arguments are, if I am right, unsound, they help us focus on some important questions, which can be taken forward into the rest of this discussion of what counts as part of the mind's physical substrate. Firstly, the phenomenal: What must be held constant for the same experience to happen, remembering that we shouldn't assume that phenomenal difference is automatically given in awareness? Secondly, the intentional: What gives a mental state its content, keeping in mind that we have to account for cases of non-veridical but nevertheless contentful perceptions? Thirdly, the causal: How do we distinguish, in the case of physically realised, conscious mental states, between what causes the state and what is constitutive of it? I will address the last first.

One way of approaching this final question is to deploy the concept of supervenience. Earlier (§3.3 Supervenience & Realisation), we said that if A supervenes on B, then one cannot attribute B to something and withhold A from it. So the question becomes, what physical conditions have to be the case so that some mental state must be attributed to that portion of the world? This is where we must be on guard against the 'causal-constitutive error' error, that is, 'objecting that externalist explanations give a constitutive role to external factors that are 'merely causal' while assuming without independent argument or criteria that the causal constitutive disjunction coincides with some external/internal boundary' (Hurley, 2006, p. 4). What constitutes the mental state is its supervenience base, that is, what must be held constant physically for the mental properties to remain unaltered. There is no principled reason to draw the causal/constitutive boundary at the skull; some parts of the brain may be more correctly seen as causal rather than constitutive of conscious mental states, and some parts of the body and world may be constitutive rather than causal. There must be a 'border' somewhere, perhaps one that is not as fixed as a cranium. If, as sometimes happens when neuroscientists talk loosely, everything is taken to be causal, then either the mental is constituted by nothing physical, forcing a retreat to dualism, or the mental is eliminated. Neither option is appealing for reasons given earlier (§2.1 Reducing Reduction).

The example of split-brain subjects (see §6.3.4 In Two Minds) is again useful to consider in this regard. In normal brains, the two hemispheres pass information between each other via the corpus collossum, in such a way that both hemispheres form one unified consciousness. When this connective tissue is cut, subjects normally continue to function as unified subjects even though information is no longer being passed between the hemispheres 'internally,' as the information is shared 'externally,' for example by objects in the hands being visible to the visual fields of both hemispheres. If the corpus collossum is taken to be partly constitutive of the experience of the subject due to its function of coordinating the hemispheres, why not take the parts of the external world that perform the same function in the case of the split-brain subjects to be internal to their mind? Moreover, if this is plausibly the case in split-brain subjects, it might be that such 'external' information sharing channels are utilised in the normal case too.

Internalist intuitions rest on supervenience thought experiments (STEs) that hold internal factors constant while varying external factors (Hurley, 2006). The classic example is of brains in vats, which has us imagine that everything you think you are perceiving now is actually the result of inputs being fed to a disembodied brain in an evil cognitive scientist's lab. The conclusion we are led to draw is that the conscious mental states would be identical in both the real and experimental scenarios if the input is the 'same,' so mental states supervene on facts internal to the brain alone.

However, for this to be a conceivable experiment, explanatory separability is necessary, that is, we have to be able to imagine 'unplugging' the brain from the world. If internal and external factors are not 'unpluggable,' but vary together, then there are no valid STEs of the brain-in-vat variety. In complex, non-linear, dynamical systems non-separability is common (Hurley, 2006, p. 7), undermining the sense in which certain factors causally explain the system's behaviour while others are merely background conditions, particularly when we see ourselves, evolutionarily and developmentally, as part of such a dynamic (§5.2.1 Evolution). Furthermore, even if some sort of 'narrow' mental content does not vary if the relevant internal factors are duplicated across different environments, meaning that a certain kind of content supervenes on those internal factors, this is not sufficient, only necessary, for those internal factors to be reductively explanatory of the 'broad' mental state as involved in worldly action. This is due to the impossibility of keeping the internal factors constant while varying the external factors, as required for a controlled STE.

Even if we could somehow set up a controlled experiment along these lines, we would have to be cautious about drawing conclusions about brains outside the lab, since we would be building a nomological machine *a la* Cartwright (see §3.1 Causation). We could set up an experiment with a brain in a vat, thus shielding it from interfering influences, and perform experiments on it, but this would only tell you some things about isolated brains in vats. Not that this would be without interest, just that it would not warrant general conclusions about actual, embodied and embedded brains in 'the wild.'

Given this, there are two ways for internalism to fail: because, granting internal supervenience, external factors are needed to explain intuition about content (content externalism), or, because mental states cannot be 'unplugged' and 'replugged,' that is, they supervene on more than internal factors (extendedness). In a dynamic system, where the agent is involved in multiple feedback loops with the environment through constant 'probing and sampling,' then unplugging will not be possible, and in order to explain the agent's actions it will be necessary to include some of the environment; the internally simulated portion will not be sufficient (Hurley, 2006, p. 141). This extension of the mental state in space is due also to its extension in time: mental states as emergent parts of a physical, dynamic system are necessarily diachronic (see §§2.2.1 Causal Closure & 3.4 Emergence). If the environment 'scaffolds' the mental state, meaning it couldn't occur without some part of the world to play its part in creating the conditions for that state, then those parts of the world are part of the supervenience base of the mental state: the mind cannot be 'unplugged' from the world it is an ongoing dynamic relationship and remain the same internally.

In response an internalist might claim that scaffolding is a supporting role and that such extended dynamics are causal rather than constitutive, part of acquiring the capacity to have full, 'quality enabling' mental states (Hurley, 2006, p. 142). However, this would be to make the 'causal-constitutive error' error: it should not just be assumed that such extended 'tuning and maintenance' processes are not an essential part of the functioning and subjective quality of experiential states.

Internalists might answer by asking about cases of non-veridical experience, where the external portion is not present, as in illusions and dreams. The thought is, if we can experience a glass both when there is no glass and when there is a glass, and it makes sense to categorise those experiences together in terms of the subjective quality of the experience, then the quality of experiencing a glass cannot depend on there being a glass there. But, given the multiple-realizability of mental states, it's conceivable that in veridical cases the realisation base of the experience is extended, whereas in non-veridical cases it is not. In fact, it is plausible that extended cases explain how we come to have the ability to simulate the experiences internally at all. Internal simulation of experience is partial and reliant on the full version that includes the world as experienced: 'What the world we are interacting with is like can be part of what enables us to experience what it is like' (Hurley, 2006, p. 146). (See §3.3 Supervenience & Realisation)

If we distinguish the cases of veridical and non-veridical experiences which are nevertheless experiences with the same content (they both tell us there is a glass in front of us), this is a kind of disjunctivism. This might seem irrelevant to the case of mental causation, since it doesn't matter whether what the mental state is about is actually the case, just whether believing it to be the case is what matters when it comes to causing actions. But, there is a distinction between cases where having a belief depends in part in on-going interaction with the world, and cases where the part of the world the mental state is about is too far away to be part of a feedback loop with the experiencer. In the latter case we can say the object of the experience is genuinely a cause of, rather than constitutive of, the mental state that is about it. In the case of the physical world we can act on, however, since embodied, extended accounts are also externalist, we do have to consider the state of the world outside the head, if that part of the world is 'in the loop,' in that direct feedback from the object is part of a spatio-temporally extended action process.

Wilson (2002) provides some reasons for scepticism regarding the embodied and extended approaches to cognition. She accuses them of not respecting our intuitions about the nature of concepts, and of not providing science with a tractable subject matter. In a version of Fodor and Pylyshyn's (1988) attack on connectionism, her first point is that concepts, such as those involved in stating the beliefs of a subject in order to explain an action, should be recombinable in infinitely

many novel ways, but that if concepts are rooted in contexts, which are necessarily finite, they cannot be.

However, as we saw (in §5.1 Rational agency), the view of concepts as fully generalisable is an idealisation. Some, rarefied and perhaps rare, concepts that have been fully abstracted from their contextual roots may meet this condition, but generality can come in degrees. This generalisability is not something that comes automatically with a concept, but is something we have to work at to achieve. A concept is not conferred generality from the moment of ostensive baptism, as the first usage of a term will have its own contextual baggage (see §1.3 Rigidity). The work involved in making a concept general is both conceptual and empirical, and the fruits of this labour may then be passed down to others.

Wilson's second, and related, point is that the objects of science are 'obligate' systems (ones that retain their identity over time) and not 'facultative' ones (that are one-off and soon dissolve as parts change) (Wilson, 2002). Since science explains obligate systems, the cognitive processes science is interested in should be general and not tied to specific, temporary configurations. She contends that distributed cognitive systems will be facultative as they contain aspects of the environment in an arrangement that will never be repeated. Such systems, she says, have vague boundaries that can only be drawn relative to the explanatory interests of the observer.

As with the previous criticism, this seems to apply an idealised standard for objects suitable for scientific discourse, one that the objects of many sciences (e.g. biology, meteorology) would fail to reach. If generality comes in degrees, and if the boundaries of objects referred to by the concepts deployed by a science are permitted to be vague and interest relative (e.g. in the study of ecological systems), then this is not a fatal flaw of a science of extended cognition; rather it is just the kind of problem that all sciences face. Her criticism would only be a problem for a strict essentialist theory of natural kinds that insists on sharp boundaries, but not for one that allows for some vagueness and observer dependence (see §1.2 Are Natural Kinds Found or Made?). Described in physical terms, mental states will be overly facultative, since mental kinds are radically multiply-realizable, and as such they will not form projectable kinds. But described intentionally they will be reasonably obligate, at least enough for us to use in relatively successful explanations and predictions. We can still use the belief that there is a glass of water is in front of me to explain my picking it up when thirsty, even though that belief may supervene on an extended portion of the world including my brain, body and the glass itself.

It is worth mentioning Rockwell (2007) as someone who takes a position that supports the one defended here, claiming it is a mistake to assume that the brain embodies the mind. This mistake is caused, according to Rockwell, by the fact that neuroscience finds the 'pragmatic causes' of mental phenomena in the brain, and assumes that the brain thus embodies the mind. By pragmatic cause he means the immediate, synchronic physical cause, but he questions that this justifies the drawing of the supervenience boundary at the surface of the brain. He argues that not only intentional thoughts, but also feelings and sensations, must be seen as supervening on the entire 'brain-body-world nexus.' We feel what we feel because of impingements on our bodies together with personal histories, and these have as much right to be called embodiment as brain cells do. He says brain activity is necessary but not sufficient for every mental state and that the borders of the supervenience base are a function of the goals and purposes of the various sciences, the distinction between the intrinsic and the relational being context dependent, thus varying from science to science.

We use our various sciences to predict the actions of individuals, for which we need to understand the social structures they are part of. These structures may be partly internalised in the individual, thus giving us the 'pragmatic' cause of behaviour, but these internalised portions may be insufficient to fully explain the action, as it may be that the agent is not fully aware of the whole structure they are a small part of, and the internal triggers may rely on the reliable presence of cues in the environment in order to bring about the 'correct' action. To quote Bourdieu (2004, p. 58): 'By constructing the objective structure of the distribution of the properties attached to individuals or institutions, one acquires an instrument of forecasting the probable behaviours of agents occupying different positions within that distribution.' A good example is the social structure that is science, and the individual scientists within in. This is not to 'relativise' scientific reason in the way that certain sociologists of science of the so-called 'strong programme' may want to: reason can be saved without transcendental arguments,

> …by describing the gradual emergence of universes in which in order to be 'right,' one has to put forward reasons, demonstrations recognised as consistent, and in which the logic of power relations and struggles of interest is regulated in such a way that the 'force of the best argument' (as Habermas puts it) has a reasonable chance of winning. (Bourdieu, 2004, p. 82)

This is cultural evolution in action, where ideas are adaptive given that they survive the selection pressures of peer review, Q&A sessions, or *viva voce* examination: 'Objectivity is an intersubjective product of the scientific field' (Bourdieu, 2004, p. 83).

Chapter 6: Physically Embodied Minds

6.1 Virtual Machines

In the previous chapters, the ground has been prepared for laying out a position on the relationship between the physical and mental that is not vulnerable to Kim-style reductionistic arguments. Now we turn to describing that position, namely Virtual Machine Functionalism (VMF). Due to constraints of space, this is not intended as a comprehensive account of that position, but rather a brief statement of why VMF is a position that satisfies the desiderata of being physicalistic without being reductionistic, or non-reductionistic without being dualistic.

VMF is a version of functionalism not susceptible to reductive arguments that use supervenience and causal arguments, because it is not just a 'black box' theory, which standard functionalism tends to be. This is because, to count as a cognitive system of the same kind as us, it is not sufficient for us to talk about the property of having the property of transferring particular inputs into particular outputs, nor to talk about fulfilling a causal role without any specification of how that role is fulfilled. Rather, the output has to be achieved in a certain way, in the case of human cognitive kinds, the way we do it, for example through the mental manipulation of intentional states. That is because, unlike atomic state functionalism (Block, 1980), VMF does not simply say that non-physical states supervene on physical states, leaving the relationship between the two levels a purely conceptual necessary identity condition. Rather, by 'state' in VMF we are referring to a state that a machine can be in; a state that is complex, holistic and diachronic, as machines are made of many parts that work together in processes with purposes, for which they have been put together, and as a result of which they have real causal powers about which we can make generalisations to be used in explanation and prediction.

Virtual machines, then, are very real, being virtual in that they are 'machines created mainly by programs running on other machines' (Sloman, 2013). Moreover, such machines could not exist in any other form, since their operation requires a speed and flexibility of function given finite resources that necessitates the components being able to be created, rearranged, modified, discarded and replaced in a way not possible with purpose built physical machines (Sloman, 2013). Being real, they are physical, therefore their causal powers are non-mysterious:

> Virtual Machine Functionalism (VMF) attempts to account for the nature and causal powers of mental mechanisms and the states and processes they produce, by showing how the powers, states and processes depend on and can be explained by complex running virtual machines that are made up of interacting concurrently active (but not necessarily synchronised) chunks of virtual machinery which not only interact with one another and with their physical substrates (which may be partly shared, and also frequently modified by garbage collection, metabolism, or whatever) but can also concurrently interact with and refer to various things in the immediate and remote environment (via

> sensory/motor channels, and possible future technologies also). I.e. virtual machinery can include mechanisms that create and manipulate semantic content, not only syntactic structures or bit patterns as digital virtual machines do. (Sloman, 2013)

An implication is that the so-called Turing 'test' is misconceived (not by Turing himself I should add). Answering questions in a way that is sufficiently similar to a human to fool a human is no criterion for the presence of human-like intelligence. Let's assume this could be achieved by 'brute force,' using computational power to search through a look-up table of all recorded exchanges. What would be missing here is any kind of semantic processing of the type we engage in. For the same reason we don't say that a chess playing machine is intelligent in any sense. This is also because it cannot take that ability and use it flexibly in other contexts. As I will argue below (§6.2 Consciousness), what gives us this ability may be the reflexivity granted by being conscious.

In virtual machines, where a machine is 'a complex whole made of interacting components,' what is manipulated is information, in the form of, for example, belief-like states about the environment that allow the machine to control its behaviour in appropriate ways given its goals. The rules governing the interaction of these components will not be the same as the rules governing the interaction of physical entities (Sloman & Chrisley, 2003, p. 145), because they are mechanisms instantiated by physical systems which are not themselves best described in physical terms, but which do support counter-factual causal claims (Sloman, et al., 2003, p. 11). Each state of a virtual machine is defined by its causal relations to other states of the system and to the environment (making them externalist, see §5.3 Externalism). These causal relations may be probabilistic, given that it cannot be shielded from perturbations from 'below and about,' which is why Cartwright's conception of causation (§3.1 Causation), where causal statements refer to dispositions to behave in certain ways given a context, is the most suitable for modelling mental causation.

Key to the operation of the kinds of virtual machines that make up the kinds of minds like ours is that they are designed to develop 'an ontology for describing its sensory contents,' based on previous experiences, by virtue of which it could 'discover within itself something like what philosophers have called "qualia"' (Sloman, 2007). Conscious, phenomenal experience, then, is not something that happens automatically through a simple activation of sensory 'surfaces,' or a taking in of information from the environment and processing it in order to produce an appropriate output. Rather, 'qualia' is the result of complex processing including the application of some kind of knowledge to the incoming information provided through the sensory channels.

This will undoubtedly seem as unsatisfactory as any other physical explanation of qualia to many. Why, after all, does that process feel like this? However, the account in question does seem to accord with how we experience the world, even if it may not always match how we think we

experience the world. However, as mentioned in relation to several points made earlier, our experience of the world might not be as we naïvely assume, that is, the properties of our experiences may be impenetrable to various, and variable, degrees. This seems problematic from the naïve standpoint because, in line with a simplistic atomic state functionalist view, the experiential is taken to be an intrinsic property of certain processes, and since nothing can explain why that kind of process has this kind of intrinsic, felt quality, it seems inextricably mysterious. It can be doubted, however, that experiential properties are in fact intrinsic, rather than relational (see §6.3.5.1 In a relationship). This is the phenomenal elephant in the Chinese Room that needs to be tamed before showing how VMF can help us find our way out.

6.2 Consciousness

Unlike standard computationalist approaches to cognition, which rely on algorithmic manipulation of symbolic representations according to syntactic properties, the causal properties of which are given by their 'shape' (physical instantiation), the approach being advocated in the present work gives the qualitative aspects of experience a central role to play, as the kinds of mental states being considered as causally efficacious parts of the cosmic machine whose patterns are visible to scientific enquiry, have a necessarily experiential aspect to them, as argued in the previous section. As a corollary, this approach to cognition also renders philosophical zombies inconceivable. In this section, I will defend this realism about qualia in a way that accords with the physicalism I have been defending, and show how this account avoids the pitfalls of other attempts at bridging the difficulties of the material/experiential divide.

If everything real is physical (§2.2 Physicalism), since phenomenal properties are real, they are physical. Most thought experiments that push us towards concluding that the phenomenal is something extra-physical beg the question by building in dualistic assumptions (see the discussion of the knowledge argument in §4.1 Mental Kinds). The dualistic intuition, that the experiential is different in kind to the physical, behind these assumptions is so strong and seemingly natural that it can be hard to shift. But, we do also have strong intuitions to the effect that our subjective, felt experiences actually cause things, at least sometimes. If the feeling itself plays no causal role, i.e. if it is epiphenomenal, then it seems hard to understand why it is universally present (if it is).

One way of arguing for phenomenal properties being causal properties is to appeal to evolution. If they were mere epiphenomenal side effects with no influence on the course of events, then why do we all seem to have them? (I'm ignoring for the moment the eliminative option that they are a

'grand illusion' and should be eliminated from our theory of mind – see §6.2.1 The Feeling of Things.) They could be 'spandrels' (Gould & Lewontin, 1979), that is, the inevitable results of our architecture (Sloman & Chrisley, 2003; Sloman, et al., 2003), without actually being selected for. But this doesn't mean they are not causal; many (most?) traits start off as 'unintended consequences' of design, but then find a role. If phenomenal properties result from the interplay of information actively gathered through navigating the world and the accumulations of previous experience, and those properties are 'visible' to selection, then they must be part of the causes of our adaptive actions. This is verging on being a viciously circular argument (they are causal if evolution can see them; evolution can see them if they are causal), but it's one that can be escaped from if we can tell a convincing 'just so' story for the evolutionary emergence of phenomenal properties by explaining what advantage having them confers. This would turn the argument from a logical circle into an emergent spiral.

One possibility is that first-person, phenomenal experience is necessary for the kind of reflexivity needed for deliberative reasoning in complex and novel environments. Moreover, being social and relying on cultural inheritance to establish adaptive 'meme lines' gives group-level adaptive function to the ability to reflect on one's experience. The ability to make a plan, and the ability to control one's behaviour and not be purely reactive, are all examples of selective pressures for organisms like us (Sloman & Chrisley, 2003). The storing of partially conceptualised representations of objects and properties, apart from allowing us to predict what the next experience will be (and thereby play a part in forming it (see §5.2.2 Expectations)), also allows 'off-line' reasoning processes about those things. It seems plausible that this kind of reflection plays a role in the formation of a self-image, something of vital importance to social animals which need to understand their role in a partly cooperative, partly competitive group, this being something that cannot be innately given, but which emerges as the result of many interacting factors.

These are speculations that will not be developed here, their aim being to make it at least plausible that phenomenal properties, including those that accompany belief-like and desire-like states, are visible to selective forces. Before continuing, let me reiterate that, in this picture, phenomenal experience outstrips introspective awareness: our introspections can be mistaken, not just about what experience is representing as being the case, but also about what we take it to be representing. That is, as well as being wrong about the dress being gold, I can be wrong about experiencing the dress *as* gold. Perhaps this is most clearly seen in the case of 'blindsight,' where a person who lacks conscious visual awareness can nevertheless perform certain physical tasks to a level that demonstrates that some visual information is available to use (Milner & Goodale, 1995).

Another example is colour constancy illusions (Adelson, 1995), where a patch of colour can be experienced in different ways depending on its context, even though in terms of the 'narrow' experience of patches of colour, the same colour experience is being had.

In this version of the 'checkershadow' illusion, the 'black' and 'white' squares seen above and below the wooden block, are actually the same grey, as can be seen when they are joined by a line of the same colour. https://www.itp.uni-hannover.de/~zawischa/ITP/introcol.html

We can be more or less aware of aspects of our visual experience. Sometimes, we consciously think about the objects around us and may form propositions as part of the act of thinking through an action. At other times, our actions are guided by a more practical kind of knowledge of the world, the exercising of mastery and imagination. This doesn't mean that it is false to describe such actions propositionally. Science describes aspects of the world using language, and such statements are true if the terms in them refer to aspects of the world that can be said to fall under that category by sharing similar enough relations to other parts of the world. Dogs don't form propositions about bones, but it can nevertheless be true of a dog that it thinks the bone is buried behind the tree (Chrisley, 1995).

In the following section, I will present an argument against an argument against a position I do not support: namely panpsychism. This is a case of 'my enemy's enemy,' our common foe being epiphenomenalism:

> Full recognition of the reality of experience… is the obligatory starting point for any remotely realistic version of physicalism…. It follows that real physicalism can have nothing to do with physicSalism…. unless it is supposed – obviously falsely – that the terms of physics can fully capture the nature or essence of experience….[W]e have no good reason to think that we know anything about the physical that gives us any reason to find any problem in the idea that experiential phenomena are physical phenomena. (Strawson, 2006, p. 4)

The attack I will defend panpsychism against, the combination problem, however, will ultimately undermine the argument for panpsychism in another way, that is, through making conceivable the emergence of the phenomenal aspects of mental states from non-phenomenal parts.

6.2.1 The Feeling of Things

The question 'What is consciousness?' can be given the straightforward (Louis) Armstrong-Block answer: 'If you gotta ask, you ain't never gonna get to know' (Block, 1978, p. 281). This reply captures the spirit of the immediate reaction to a philosopher questioning consciousness. It's right there, transparent to me, nothing about it seems opaque; one thing that I know more intimately than anything else is what it is like to be me. This is part of the Cartesian certainty, used not only as a basis for our understanding of the mental, but of mathematics and science. I clearly and distinctly perceive a patch of colour, and I cannot doubt that I am perceiving *that* colour.

We should take Descartes' method of doubt further than even he did and question the seeming clarity and distinctness of our 'inner' experiences (it may be true that I can't doubt that I am thinking, but it can be doubted that I know what I'm thinking). When we refer to these private experiences, how can we do so without using terms the meanings of which rely on publicly accessible meaning? I'm not denying that we have experiences, obviously, just saying it might be fruitful to apply a measure of methodological scepticism to what seems to be a given: 'There seems to be phenomenology, but it does not follow from this universally attested fact that there really is phenomenology' (Dennett, 1991, p. 365).

Strawson accuses Dennett of being 'so in thrall to the fundamental intuition of dualism… that [he is] prepared to deny the existence of experience' (Strawson, 2006, p. 5). That is, Dennett is accused of assuming that since experience is of such a different nature to the material, and since everything real is material, then the experiential cannot be real. Strawson, on the other hand, says, 'For there to seem to be rich phenomenology or experience just is for there to be such phenomenology or experience' (Strawson, 2006, p. 9). But, as we will see, Strawson is himself in thrall to dualistic intuitions, as he relies on the 'never-to-be-reconciled' distinction between the experiential and the material in his argument against emergence, which is a major premise in his argument for panpsychism.

Nagel gave voice to the common intuition: 'It is difficult to imagine how a chain of explanatory inference could ever get from the mental states of whole animals back to the proto-mental properties of dead matter' (Nagel, 1979, p. 194). Despite this, we also have the intuition that if conscious mental states have physical effects, then they must, in some sense, be physical. Is it that our intuitions regarding the physical are incorrect or incomplete, or our intuitions regarding the experiential? It will be my conclusion that the experiential resides in 'virtual' processes (in the sense discussed in §6.1 Virtual Machines) instantiated in but not reducible to the material.

Aleksander & Dunmall (2003, pp. 9-10) offer a set of axioms that give minimal necessary conditions for the presence of consciousness. These are:

1. Perception: having depictions of the world.
2. Imagination: recalling such states and fabricating states that resemble them, the intentionality of which is 'inherited' from experience of the thing 'out there' that they are derived from.
3. Attention: being able to direct attention to certain aspects of the world, this being made necessary by limited bandwidth of perceptual channels.
4. Planning: being able to use imagination to contemplate and plan actions.
5. Emotions: having affective states that evaluate and motivate/veto actions.

Now, the question is whether the kinds of mechanisms that need to be present for the agent to have a sense of its own presence in the world, as well as the affordances for action open to it, are also sufficient for consciousness. Aleksander & Dunmall say we can assume a one-to-one mapping between sensations of consciousness and the neural mechanisms that underpin the functions set out in the axioms, but that the science of consciousness should only concern itself with this functional description, leaving the subjective sensations alone (Aleksander & Dunmall, 2003, p. 15). Notice that the identity relationship assumed here is not between mental and physical kinds, but mechanisms described functionally and mental kinds. Similarly, Franklin (2003) says a creature has 'functional consciousness' if it possesses states and mechanisms that make it aware of dangers and opportunities in the world and allows it to navigate and manipulate the world in accordance with its needs, but distinguishes this from phenomenal and self-consciousness.

Can we separate and scientifically side-line the subjective quality of experience, or is this a necessary part of the causal story of the world? Global workspace theory (Baars, 1988) claims that consciousness is what happens when otherwise unconscious, specialised cognitive processes 'form a coalition' to solve a particular problem, possibly recruiting other processes to the task as appropriate. There is competition for attention, but unopposed ideas will be acted on (cf. James' (1890) ideomotor theory). This would give a special causal role to consciousness, if it is seen as necessary for performing this function. But what is it about that kind of function that merits a phenomenal rather than a functional description when it comes to scientifically analysing the causal processes?

This question is of the same form as the one discussed earlier (§2.1.1 The 'Special' Debate) about the possibility of taking physically realised mental states to be causal, and therefore as having downward

causal influence despite supervening on the physical states. There (and later in §3.4 Emergence) it was argued that physicalism does not rule out emergent causal phenomena, within the constraints of the physical. Furthermore, talking of the causal powers of mental states in purely functional terms does not exclude an explanation of how those causal roles are actually cashed out, in this case in terms of the functional role of consciousness. The aversion to cashing out functional talk in these terms seems to me to be something like a kind of scientific taboo at mentioning the 'c' word, which I think is a result of implicitly accepting the 'mysterian' view that the relationship between conscious states and physical ones is intractable (McGinn, 1993), and therefore the subject is best avoided by 'serious' scientists. However, it seems to me that if we can tell a coherent, natural story of how conscious states come to be, and what they are for, then we should not be embarrassed to talk about it. Indeed, if, as I am arguing, we can't make sense of the causal properties of mental states without referring to the perceiving subject that has them, then we are obliged to include them in our scientific discussions.

Opposed to the mysterians, are those who do not find it inconceivable that the relationship between the phenomenally conscious aspects of the mind and the physical states that realise them should be beyond our ken:

> …every phenomenal kind M is identical to some P that is generally similar to the kinds currently recognized by the physical sciences…. [W]hen we have established such M=P identities, then we will therewith have 'fully captured the nature or essence of experience' in physical terms, in that the relevant physical term will refer to nothing other than the phenomenal kind M. (Papineau, 2006, p. 101)

However, this might be too strong, if it is interpreted as implying a type-identity (see §3.3 Supervenience & Realisation). If VMF is right, there may not be such identities, as the explanatory work will be done by the 'design' of the machine, with the physical components being necessary but not sufficient for explaining the causal properties of the VM.

If there were some intrinsic properties of some physical kinds that explain consciousness, and if our brains are the seat of our consciousness, then there is some intrinsic property of neurons that is a conscious property. But neurons are not individually conscious; it is only when they are part of the system that is the brain (plus body and world), that they become part of a pattern that has phenomenal properties. Thinking that neurons contain a little bit of consciousness due to their intrinsic properties is akin to the panpsychist thought that everything does, and suffers from the same problems. Rather than being an identity between mental and physical kinds, consciousness is better seen as the result of a certain kind of information processing. For this reason, I can conceive of a robot as having consciousness if it is suitably hooked up to the world and processes information

about the world and its own internal states in the right way, but I cannot conceive of a rock having consciousness, as I don't think it processes information about its environment, or monitors its own state.

Our form of embodiment partly determines how the world appears to us, through directing our attention to aspects of the world relevant to us, meaning the information from the environment is to an extent 'pre-processed.' A 'pure' information processing understanding of mental states sometimes advocated by traditional computational functionalism, for example the 'sense-model-plan-act' framework, is too abstract and insufficiently dynamic. It would be computationally intractable if all the information needed processing so as to select the salient patterns, and from this a detailed model of the world had to be constructed for use in planning actions. An embodied approach means we don't need to solve the processing and model building problems, as the information is presented to us in a structured, meaningful way. Cognition is 'scaffolded' by the socio-cultural environment, which has an evolutionary dynamic and which we have evolved to internalise as part of our development. For this reason, embodiment theory takes phenomenology seriously: how you perceive yourself and how you perceive are related in a self-sustaining feedback dynamic, and an understanding of this first-person perspective, of how the world is presented to subjects, is an integral part of the scientific project of understanding mental states and their causal propensities.

One problem with this discussion is that old, disembodied ways of thinking are embedded in the language, so deeply that it affects even the prepositions we use. Our experience is not something *of* which we are aware, rather it is something *with* which we are aware of the world (Rowlands, 2002; Clark, 2002). The mistake of thinking that what we are directly aware of is our experience leads to the Grand Illusion Illusion: we are led to scepticism about our experience because we know our phenomenal experiences can be misleading, but that scepticism is misplaced: rather than our perception seeming to be a particular way, it is the world that seems to be a particular way according to our perceptual mechanisms that have developed to do their best with the information at their disposal. So, although we may not have full colour experience at the edges of peripheral vision, despite the fact that it seems that we do, this does not mean that we are mistaken; the things in your peripheral vision seem to have colour, and indeed they do. We experience them as coloured because of expectations we have developed through experience (see §5.2.2 Expectations).

Of course, under certain circumstance our expectations can lead us astray and then we would be in a state of illusion, but not a grand one, just a normal one, where we are misled by out-of-the-ordinary situations. Our expectations are based on what has gone before and the generalisations we have

formed as a result of recognising patterns and abstracting from experience. This is a process that takes place over (at least) three different timescales: phylogenetic (we developed the perceptual mechanisms to make sense of the world we find ourselves in); ontogenetic (we build up a lifetime of experiences); pragmatic (visual saccades build up a picture of our immediate environment). It is not surprising that these processes have produced mechanisms that are not always accurate, after all, they are evolutionary, and so imperfect in the materials at hand to build them, as well as needing to be economical (§5.2.1 Evolution).

The Grand Illusion hypothesis finds support in much work on perception in recent years, e.g. change blindness. It's a form of the classical sceptical argument: given that we are wrong about some perceptions, we might be wrong about all of them, and is a symptom of the disembodied approach to cognition that places too great an emphasis on internal representations rather than seeing perception as intimately connected to the possibilities of action in the world (O'Regan & Noë, 2001). However, although there is merit in the enactive approach, it ties perception too closely to immediate action routines and fails to see the much longer term processes involved in building up perceptual routines (Clark, 2002). So, the embodied account of perception needs supplementing with the idea of cognitive mechanisms that function to 'fill in' the relatively sparse information available at any moment from the outside world. This is a top-down process of feedback from experiences, which is fed forward into action, in conjunction with immediately available information on sensorimotor contingencies (§5.2 Feedback and Feedforward). 'Radical' enactivists seem to be forced to concede that any device that uses information in the formation of actions would count as an experiencer, including a thermostat, unless they want to add architectural conditions to rule out such devices, in which case they are not radical anymore.

Another example where subjective phenomenal experience and physical action seem to be mismatched is the Ebbinghaus illusion, where two circles of the same size appear to be different sizes because of the relative size of a surrounding circle of circles. Although the inner circles appear to be different, this information does not result in a difference in motor behaviour when reaching to grasp the circles (Goodale, et al., 1991). On a more everyday level, our rapid reactions to things in the world, like small objects moving towards our eyes, happen too quickly to pass through higher-level processing and decision making procedures. In Goodale and Milner's (1992) account, there are two separate neural pathways that process information from vision. One is the older system for guiding skilful action, the other is newer, for object recognition based on knowledge and memory.

This newer system is used to classify things seen in the world into kinds based on our conceptual schemes, which may be updated through receiving new information, and which could be used for

the deliberative forming of decisions to act. Here our first-person perception of the world comes together with the scientific project of describing kinds in the third-person in order to explain and predict. Feedback from these higher centres may be necessary for visual awareness of certain types and certainly affect the contents of perception, and may be involved in action selection (Pascual-Leone & Walsh, 2001). This is a different sort of action than that resulting more automatically and fluently in our habitual actions, like that of walking. The difference between these two mechanisms of action is shown by the interference that can happen when deliberative thought is applied to habitual actions; think of the feeling of awkwardness and lack of coordination that results from becoming aware that you are being watched while walking, and consciously trying to walk normally.

There are different terms of art used in the literature to refer to conscious experience, which I have been using without much discrimination so far, e.g. 'qualia' (Lewis, 1929; Nagel, 1974) or 'phenomenal consciousness' (Block, 1978). These bring with them assumptions about the essence of experiential states, for example that they are intrinsic, ineffable and incorrigible etc. These assumptions may be a source of theoretical problems rather than statements of truths we should automatically assent to. This doesn't mean that we should assume the opposite, that we are misled in thinking that conscious decisions are real, or that they are stories we tell ourselves to maintain a fiction of ourselves as a unified subject (Dennett, 1992). We can have veridical experiences of ourselves acting on our deliberations without assuming properties like intrinsicality, etc.

The intuitions captured by the philosophical concept of qualia are, however, deep seated and not easily dispelled. The feeling that there is a gulf between physical and phenomenal explanations is not bridged by statements of identity, even if true. Let's say that it has been adequately empirically shown that the physical correlate of pain is C-fibre firing. Some, e.g. Papineau (2006), would say that such an identity is an *a posteriori* necessity just like 'water = H_2O,' and that the intuition of an 'explanatory gap' (Levine, 1983) is just an illusion. The nature of water is not given to us directly *a priori*, and neither is the nature of pain; both refer to what they do by virtue of certain causal and historical facts (Papineau, 2006, p. 102). The illusion of the gap stems from confusing merely 'mentioning' a conscious state and 'using it.' When we make an identity statement between a phenomenal concept and its physical correlate, we are just referring to kinds of states in the world that we have discovered are identical in extension; the phenomenal concept doesn't 'capture' some deeper essence: 'Scientific talk of relevant brain states picks out states which are in fact essentially conscious, but does not a priori display those states as conscious' (Papineau, 2006, p. 106). When mentioning a mental state with phenomenal language, 'the conscious referent seems to be present in the thinking itself,' and this apparent transparency seems to give us direct access to the essential

properties of that mental state, which don't include the physical properties, leading us to mistakenly conclude that the physical state is not identical to the mental one. However, phenomenal concepts, just like other intentional concepts, 'gain their referential powers from causal and historical relations, and those referential relations can leave many essential features of the referents opaque' (Papineau, 2006, p. 105).

Papineau argues for physicalism on a causal-explanatory basis, and says the 'brute intuition' that 'straightforward' physicalism cannot be true is just an illusion that would continue even if it were proved to be false. He equates the intuition to that of the earth standing still, in that even when we understand inertial forces we still feel ourselves to be standing on a stationary surface. I think there is a difference, though, which is that we can make sense of the feeling of being still given the laws of physics, whereas the anti-physicalist intuition stems from the fact that knowing the physical laws 'underneath' experience doesn't seem to explain what it feels like to be in that physical state. The problem is not that the concept of a physical state excludes consciousness, but that it doesn't seem to necessitate it.

It seems 'completely open what it would feel like to be a purely physical being with firing C-fibres…. Why suppose it must feel like nothing?' (Papineau, 2006, p. 101). As mentioned, Papineau relies on the argument that many necessary identities are established after the fact, like 'salt' and 'sodium chloride.' However, earlier we saw reasons to doubt a straightforward causal theory of reference that assumes we can label parts of the world without using assumptions about the nature of those parts, leaving us free to investigate those natures later when we have the conceptual tools to do so (see §§1.2 Are Natural Kinds Found or Made?, 1.3 Rigidity). But, if baptisms are not as conceptually blank as pointing, and require prejudices about the nature of the parts being pointed at in order to actually succeed in referring, then it might not be as simple as just investigating *a posteriori* what physical parts of the world are identical with the mental parts we have picked out with phenomenal concepts. If so, we couldn't simply 'allow that the term C-fibres firing fully captures the nature and essence of Pain' in the same way that 'the term sodium chloride fully capture the nature and essence of table salt' (Papineau, 2006, p. 102). This is because we cannot assume that all natural kinds have micro-physical essences that we can point to with a name, said name being a placeholder for whatever it is that micro-physically determines all the other properties associated with that kind of thing. Part of the essence of some things is determined by how they are related to other stuff in their environment, by the role they play in wider processes, roles that might be realised by virtual machines.

McGinn (2006) wonders if Strawson's assertion that experience 'just is' physical is substantive, or is just a synonym for 'concrete,' as Strawson doesn't say in what respect it is physical, for example, being spatial, causally connected to non-experiential facts, or something you can bang your head on: 'We need to be told in what respect experiences are like molecules before we can assess whether the class constitutes a genuine natural kind…. He simply wants to call experiences physical – just as I may want to call ocean waves spiritual' (McGinn, 2006, p. 91). This seems a little unfair, unless McGinn is saying this in light of an ontology that claims all real things are spiritual and waves are real, and then he would need to back up this spiritualist position, but he wouldn't have to say in what respect waves were spiritual.

McGinn's second objection is that Strawson's use of the term 'physical' is a 'flagrant violation of common usage' (McGinn, 2006, p. 91). But can we say that a philosophical term has a 'common' usage? Moreover, if we think that there is a fundamental misunderstanding entwined in the common understanding of a term, then we should be able to challenge that usage. This is in effect what Strawson is doing, by taking physicalism as saying that all concrete phenomena are physical, since by common understanding, many things are not physical, for example, feelings. McGinn claims that physicalists believe that 'experiential facts all supervene on non-experiential facts…, and that the causal powers of [experiential] facts…, are specifiable in entirely non-[experiential] terms' (McGinn, 2006, p. 90). However, some physicalists would deny this for a number of reasons: they might not accept supervenience (e.g. Cartwright (1999), Humphries (1997a)), or could deny reduction ism by saying, as I am, that the causal powers of experiential states are not specifiable in non-experiential terms.

After characterising physicalists the way he does, McGinn says, '…they seem to miss out the very essence of what an experience is' (McGinn, 2006, p. 90), and concludes that Strawson should drop the physicalist terminology, since it is just playing with words and doesn't help answer questions like how the mental and the physical are related: 'By his methods we could extend the reach of physicalism still further, by declaring that 'physical' is a natural kind term for such things as bodies, minds and numbers!' (McGinn, 2006, p. 92). Could we extend the physical to abstracta? They exist in the mind, but not as concrete things. It seems far from trivial to claim that all concrete phenomena must be at base physical. Strawson's problem is to reconcile the physicalist motivations with experiential facts that are 'totally different in kind' (McGinn, 2006, p. 91). A possibility is that this difference comes about as a result of differences in descriptive perspective, i.e. the first- and third-person descriptions of the same events. However, Strawson denies that a first-person perspective can emerge over time in physical systems, so ends up needing to assert that all things

describable in third-person terms also have some sort of first-person properties too. In the next section, we consider this argument for panpsychism.

6.3 Panpsychism & Composition

Panpsychism states that there must be some fundamental property of matter that disposes it towards constituting experiential states, and that this property must itself be, in some sense, psychical. That is not to say that each particle has conscious experience in the way that we understand consciousness from our own experience, just that there is some intrinsic property, separate from those properties described by third-person physics, that is responsible for collections of matter like us having first-person experiences. This view relies on a supervenience claim about the relationship between macro-experiential states (like we have as subjects), and micro-experiential states (seen as intrinsic properties of constituent particles). Even if it is accepted that such micro-experiential states could exist, there is a problem faced by this view, namely, it seems clear that, as subjects of experience, we are singular and unified; there seems to be one experiencer present in each of us rather than many. How are macro-experiential states supposed to emerge from their micro-experiential components? In the case of the emergence of liquidity, both the component parts and the macro-state have causal descriptions that can be explanatorily connected, but it doesn't look like micro-experiential states could explain macro-experiential states in the same way (Papineau, 2006).

6.3.1 The Living Dead

The conceivability of philosophical zombies is taken by many, e.g. Chalmers (1996), to show that experiential properties are not physical. A zombie-you, being physically and causally identical to you, could be sitting reading this, frowning, expressing frustration, but nevertheless lack any experience or feelings. The zombie lacks conscious awareness of its place in the world; it is just a bunch of cells, adding up to no more than the sum of its parts. If this is conceivable, so the argument goes, experiential properties are not physical as they are in addition to all the physical goings on of the zombie, so physicalism is false. However, an unfortunate side-effect of this is that your conscious, experienced thoughts are epiphenomenal; they are not the causes of your physical behaviour, since the zombie-you's frown was caused in just the same way as yours.

Panpsychists attempt to eliminate the gap between the material and the mental, and thus avoid epiphenomenalism, by installing experiential properties with physical ones in the building materials. As mentioned above (§2.2 Physicalism), Russell (1927) pointed out that physical science only captures the extrinsic, relational properties of matter, and thus is silent on the intrinsic nature of matter, and later said, 'We know nothing about the intrinsic quality of physical events except when these are mental events that we directly experience' (Russell, 1948). Another such position is Chalmers' (2003) panprotoexperientialism. This leads Strawson to ask, 'Why then… do so many physicalists simply assume that the physical, in itself, is an essentially and wholly non-experiential phenomenon?' (Strawson, 2006, p. 11). He claims that the idea of the non-experientiality of the physical is at odds with the thesis of realistic physicalism, that experience is a real concrete phenomenon and every real concrete phenomenon is physical (Strawson, 2006, pp. 11-12). Strawson accepts that 'experience is 'really just neurons firing… [but that] certainly doesn't mean that all characteristics of what is going on, in the case of experience, can be described by physics and neurophysiology…" (Strawson, 2006, p. 7). Given that emergence is not an option (see below), and that in the case of experience there is no real distinction between reduction and elimination, if experience can be reduced wholly to the non-experiential, then it doesn't exist.

Strawson's argument is a kind of inference to the best explanation. Given the obvious reality of experience, and the fact that everything real is physical, and given that experiential properties cannot be derived from those found in physics, then there must be properties that are not described in the language of physics. These psychical properties have either emerged somehow from a non-psychical background, or they were there from the start. Since radical kind emergence (as opposed to 'moderate,' epistemic emergence, as in that of liquidity) is untenable, then all fundamental particles must have psychical properties.

This untenability, though, is based on the same assumptions about the nature of experience that created the problem in the first place. That is, it is claimed that, since experience has these special properties, then it can't come in degrees; it's either there or not. It can't just 'pop' into existence, and it isn't necessitated by the right combination of non-experiential parts, so it can't emerge. We will return to emergence shortly, but first we should focus on the main problem for panpsychism, namely the composition problem: given the unified nature of an experience, it cannot be composed of many smaller experiences. I will claim that there is a solution to this problem, but that solution in fact gives us reason to believe that emergentism is tenable, and so emergence is the better explanation after all.

6.3.2 The Mind-Mind Problem

A problem that panpsychist theories face in placing minds like ours in a physical world is that it is not immediately clear how lots of little minds can add up to one big one. How can 'smaller' minds be related to each other in such a way as to lead to the emergence of a 'larger' unified mind? The problem can be approached from two directions: the 'bottom' or the 'top.'

The former is a metaphysical enquiry into wholes and parts. James (1890, p. 160) called it the derivation problem. Composition is unproblematic for particulars like a bath-full of water: it is a collection of water molecules the properties of which explain, together with the context, the properties of the whole. In the case of experiential properties the same cannot be said: if a feeling is composed of a hundred other feelings, then, since the feeling to be explained exists in addition to the collection of feelings that are supposed to explain it, there are 101 feelings in existence, and we still need an explanation of the extra feeling. The whole is greater than the sum of the parts, as the micro-subjects don't have the right properties to explain the properties of the macro-subjectivity in a way analogous to how facts about molecular structure explain liquidity. If a new feeling comes into existence when the 100 feelings are put together, then this is a case of 'radical' emergence, and, if that is acceptable, there is no need to posit panpsychism at all.

The 'top down' approach is based on phenomenological considerations. From our point of view, we seem to be unified subjects. There may be many different things happening in one's mind, but they are all happening to one: there is one subject having these experiences. These arguments rely on the privilege of first-person authority, but not in a problematic way; even if we doubt that we are unified subjects, we still have to explain the sense of self we do have. Some just state it as an *a priori* fact that subjects can't compose, for example the principle of 'no summing of subjects' (Goff, 2009, p. 302), which is based on a simple disanalogy between objects and subjects (objects can sum, subjects can't).

There seems to be a singular thing which is our selfhood, which does not appear fully-formed but comes to be through the coming together of parts over time, but this is at odds with the intuition that a self is neither composed of 'part-selfs' or non-experiential physical parts. But, if we can answer 'easier' questions (Chalmers, 1995) of cognition, e.g. how the visual system works, and we can take such systems to have mental properties, then we have to face the question of how these combine to form a whole self, without falling back onto implicitly inserting a homunculus to bring them all and in the mind bind them. I call this question the Mind-Mind Problem.

To explain how subjects come to be through natural, spatio-temporal processes, either these processes involve non-experiential or experiential parts; in the first case, we have to bridge the

'explanatory gap' between non-experiential descriptions of physical processes and experiential descriptions of subjective states; in the second case, states of subjects seem to be unified in a way that cannot be explained by decomposition into constituent 'semi-subjects.' The solution is to see that this seeming impossibility of decomposing subjectivity is an illusion, thus allowing that two or more 'semi-subjects' can come together to form a subject. This is not to deny the existence of selves, as some would (e.g. Metzinger (2003)), but to enhance our understanding of what subjects are. Moreover, this does not constitute support for panpsychism, but for physicalism, as it attempts make the emergence of subjectivity from non-experiential parts conceivable, and in the process bridge the explanatory gap. Before presenting the argument, we should examine the self, to be clear about what we are trying to explain.

6.3.3 Feeling Things

Seeing experiential properties as intrinsic properties of subjects leads to the 'hard-problem' of explaining why cognitive processes are accompanied by the subjective feelings that they are (in contrast to the 'easy problems' regarding the mechanics of particular cognitive processes) (Chalmers, 1995). In the following, I aim to show that the existence of answers to the 'easy' questions may add up to an explanation of everything that needs explaining.

The assumption that experience is an intrinsic property of certain brain processes means that there can be no external explanation of why a particular neural event has the felt quality it does. It follows that the extrinsic, relational descriptions of the physical sciences will never capture these intrinsic properties. Moreover, we can never be wrong about how things appear to us because these intrinsic properties are just there for us; we can't be wrong that we are experiencing the world a particular way, even if we are wrong that the world is the way we are experiencing it to be.

Both the panpsychist and unitarian (those who see the mind as a simple, uncomposable unit) accept the inconceivability of the emergence of the mental from the physical. Emergentism comes in ontological and epistemological varieties (§3.4 Emergence). Strawson accepts the latter, but denies the possibility of the former, because 'only like can emerge from like' (Strawson, 2006, p. 15). His argument against ontological emergentism is based on the thought that 'it seems plain that there must be a fundamental sense in which any emergent phenomenon… is wholly dependent on that which it emerges from' (Strawson, 2006, p. 15). For example, liquidity from collections of H_2O molecules at room temperature near the surface of earth. Of course, this assumes the mental is fundamentally unlike the physical.

Despite the fact that many take this account of the nature of phenomenal experience to be clearly and distinctly true, it can be doubted. What does it mean for these intrinsic properties to be 'just there for us'? It isn't clear to me that these 'felt properties' are purely intrinsic and immediate, and we have seen examples from cognitive science that should lead one to question assumptions as to the incorrigibility of first-person experience (§6.2 Consciousness). In the case where two patches of colour appear different but can be shown to be the same, we are forced to ask which of the experiences was intrinsic to the physical happening being caused by light from those patches hitting our retina and being processed. The light that goes to your retina from that patch of colour does not change, and neither do any of the intrinsic properties of the neural processes that are associated with perceiving just that patch of colour. But top down processes that utilise relational properties, do affect how that patch appears to us. Which one is the 'real' experience?

Rather than see consciousness as a fundamental property of matter like panpsychists, or even a property of particular types of biological matter (e.g. (Hameroff & Penrose, 1996; Searle, 1992), it is more useful, I would argue, to think of it in terms of particular kinds of processes: feeling an emotion, attending to a flash of colour, expressing ones thoughts, and so on. The mistake, it seems to me, is to think that there is an emotion, colour, or thought that exists separately from the feeling, attending, or expressing. This way of thinking fits naturally with the idea of the brain being divided into specialised, functional parts in which particular processes take place (Steels, 2003, p. 173). The question, then, is where the self is among these processes. The self can't be a 'homunculus' experiencing these processes; the processes *are* the experiences. But in that case, if there is no experiencer separate from the experiences, how do these experiences add up to make a self? This position faces similar epistemic and phenomenological problems of composition to panpsychism. The epistemic problem is that if *a* subjective state of mind is composed of other states of mind, it seems the macro-state cannot be explained merely by summing together those other states. The phenomenological problem is that I feel like a unified self, the centre of my mental world; I don't feel like a fragile and fragmented collection of separate mental states (most of the time).

My strategy in the following is to give empirical support to the idea that selves (and therefore the experiences they have) can combine to form other, 'larger' selves, but then to argue that rather than supporting panpsychism, this supports the idea that selves can emerge given the right kinds of parts connected in the right kind of ways. Since it is not mere summation of subjects (not just sticking them together), but rather connecting them in ways that mean they can function together in the real world, the derivation problem is solved. However, since it is solved by showing emergence to be conceivable, panpsychism is undermined.

6.3.4 In Two Minds

Corpus callosotomy is the procedure of cutting through the tissue that connects the cerebral hemispheres and which functions to pass information between them, effectively dividing the subject's brain in two (Gazzinga, 1970). Under normal circumstances, where the hemispheres can share information by perceiving the same world, for many split-brain subjects, everything seems more or less normal in terms of behaviour and reported phenomenology. However, when the information each hemisphere receives varies in meaningful ways, interesting phenomena arise which, I will argue (*pace* Nagel (1971, p. 409)), should lead us to say that there are two subjects present under the experimental conditions, which compose to make one subject under normal conditions. If this is the case, then we have reason to say that our intuitions about the composability of subjects should be revised, and furthermore that our intuitions regarding the nature of conscious experience, namely its intrinsicality (and the resulting ineffability, indivisibility, incorrigibility, etc.), should also be challenged.

In Gazzinga's experiments, subjects whose corpus collosa had been severed had different information presented to each hemisphere by showing different images to the visual fields of the respective hemispheres. In the most striking cases there can be conflict between the hemispheres. In a case described in Nagel (1971, pp. 400-1) the right hemisphere is shown an image of a pipe, and the left hand tasked to write the word. It struggled to do this as language capacity tends to reside in the left hemisphere. After haltingly managing a 'P' shape and the down stroke of the 'I', the left hemisphere appeared to get frustrated with the 'dumbness' it was observing, took control of the pencil, made a guess that the other half was trying to write 'PENCIL,' and finished off the word. Then the right hemisphere, seemingly irritated, took control, scribbled out the word and drew a pipe.

Many examples of such interhemispheric conflict in split-brain subjects have been documented (Wolman, 2012). Long-term studies have shown that, for some split-brain patients, each hemisphere can have has its own sensory-motor interface with the environment, as well as different memories, cognitive and linguistic abilities and repertoires, including distinct personalities and preferences (Zaidel, 1994). Split brains might also be common in nature. Birds, for example, have no corpus collosa, making them natural subjects for studying conflict between the hemispheres (Ünver & Güntürkün, 2014). In the case of conflict between the hemispheres in birds, one hemisphere takes control, but in humans we witness conflict. When a person whose hemispheres are in conflict is asked to explain his or her actions, the subject will generally rationalise events in a way that preserves his or her sense of self: a confabulation. Nagel gives the example of presenting a picture of a naked woman to the right hemisphere of a male subject, who exhibits a typical

physiological response; when asked why he is excited, the left hemisphere, interpreting this excitement in terms of what it can see, replies, 'Wow, that's quite a machine you've got there' (Nagel, 1971, p. 401).

What does this show us about the nature of subjecthood? Nagel (1971, pp. 402-3) gives us five options (although he talks in terms of 'having a mind,' I take this as equivalent to 'being a subject'):

1) Split Brain Subjects (SBSs) have one mind, constituted by events in the left hemisphere, the events in the right hemisphere being non-conscious.
2) SBSs have one mind, constituted by events in the left hemisphere, the events in the right hemisphere being conscious but not belonging to a unified subject.
3) SBSs have two minds, one of which is dumb.
4) SBSs have one mind constituted by events in both hemispheres but which is less unified than normal minds.
5) SBSs have one mind which sometimes splits into two in certain circumstances.

The case for 1) & 2), locating the mind in the left hemisphere, rests purely on the fact that we can communicate with that hemisphere and it doesn't seem to be aware of what's going on in the right. Unless we want to say that the ability to use language to communicate is an essential property of being a subject, there seems little reason to accept these, particularly since we generally accept that non-language using animals and human infants are subjects. Option 3) is problematic because SBSs often appear to be normal in non-experimental conditions, both to themselves and to others; they act like a unified subject in all the ways that usually would cause us to count something as a subject. 4) doesn't seem to describe a subject that would fit the conditions of normal subjecthood, that is, the experiences this subject has, when in the experimental conditions, is too disconnected: we would normally think of a subject as being able to compare two distinct visual impressions. Finally, 5), according the Nagel, is an *ad hoc* move that only accounts for the phenomena being investigated, neglecting the fact that there will be other experiences had by the SBS at the same time which are not separated by the machinery of the experiment and should thus count as being experienced by a single subject: there can't be both two subjects and one (Nagel, 1971, p. 408).

Nagel concludes inconclusively, saying that what is true of SBSs is also true of 'normals,' as we are made of two cooperating hemispheres too, and since there is no definitive answer as to whether SBSs have one or more minds, there is no definitive answer whether we have either. He speculates that future scientific discoveries might render the question meaningless, by showing the whole idea of a unified subject to be 'quaint,' although we may find ourselves 'unable to abandon the idea'

(Nagel, 1971, p. 411). We are in the future now, and there are people in both camps; those who reject the notion of a self as a fiction or illusion, and others who would say that what we have learnt informs our concept of self, rather than eliminating it.

The position being advocated here is that such data should change our concept of self and subjecthood without leading us to reject those terms. Our idea of the self has definitely been informed by science since Descartes' day, when it was inconceivable that the conscious self could be a physical thing (and this intuition is, as we have seen, still exerting its force). We have to let go of the idea of an idealised self as the 'central experiencer in chief.' The self is more distributed and fragmented than that, perhaps more so than we like to admit, and understanding this may be the therapeutic benefit of cognitive science. But neither is it so fragmented as to render it meaningless to talk about the self as something that has an identity in space and time, one that science can make meaningful statements about.

'Normal' selves made of closely cooperating parts are clearly useful, both for the person involved and any scientists trying to make general statements about them. Two people closely cooperating are not a single self, as they don't behave in a sufficiently unified fashion, because they don't share a mental life: if asked you could get two different answers to the question, 'Why are you doing that?' The hemispheres of some SBSs do sometimes seem to be sufficiently disunified to be thought of as two separate subjects. The most interesting cases, for my purposes, are those who only exhibit this disunity under experimental conditions, seeming normally at one with themselves the rest of the time. If we take a subject to be a 'locus of experience and action' (Nagel, 1971, p. 405), in other words, something that can take in information about the world and use it to produce intentional behaviour, then it seems that both hemispheres of split-brain subjects should be counted as separate subjects when the information they are receiving at a time is different. After all, if a person loses the left hemisphere but survives, we would treat the remaining person as a subject worthy of ethical consideration.

Is there anything more we can say about what makes a collection of parts a single subject beyond behavioural evidence, including subjective self-reporting? After all, a closely cooperating group of organisms, like ants or football fans, can act *as if* they were an intelligent agent, and subjective reports are unverifiable without further criteria of correctness. Furthermore, a football crowd might well report their feelings regarding, for example, the referee's abilities. Anyway, the ability to report on one's experience can only be a sufficient condition, so does not warrant deflating the mind to the hemisphere that can report its feelings to us (also, in Nagel's example, the right hemisphere *did* seem to express its feeling of frustration when it scribbled out the word 'PENCIL'). The productions

of my right hemisphere are normally taken to be within me; I don't have to 'listen' to it, 'read off' from it, or otherwise make an informational effort to gather its contents, for example to understand the spatial arrangement of objects around me: I will be looking at the room, not into my right hemisphere. Normally, both my hemispheres are responsible for my mental life. Although there may be a sense in which, under special circumstances, SBSs can be seen as having two separate experiencing subjects in their brains, under normal circumstances, where hemispheres share information though the corpus collosum, there is a single subject. If this is the case, when moving from experimental to 'normal' situations, two subjects become one because of the shared informational world they inhabit. Thus, I take Nagel's 5th option, and this can count as empirical support for the possibility of combination regarding subjects of experience.

6.3.5 Selfishness

The claim that two 'small' minds can become one 'larger' mind by changing how information is shared by them is a conjecture in need of more support. What makes the halves parts 'inside' the whole, rather than each of counting as 'input' to the other; what makes them constitutive of a mind, rather than each being causal with respect to the other? After all, when someone stands on my toe, a signal is sent to my brain and I feel the pain in my toe, but that kind of informational connection is not generally taken to justify counting the toe as part of the mind. What is it that makes the somewhat disparate pieces that make up a mind feel like a self, or, what makes some body selfish?

The first reason may be connected to what it is to be a representation, or to be represented. Given the above account of representations (§4.2 Representations), in which representations are states of our internal, virtual machinery that function to stand in for parts of the external world in processes where those parts are not themselves available, if the parts are causally 'close enough' for information to be exchanged with sufficient speed and fidelity, then those parts can be constitutive of a virtual machine performing a mental task, and so be part of a mind. Furthermore, as we saw above (§6.1 Virtual Machines), what makes something a virtual machine has something to do with the historical process of how the parts were engineered over evolutionary and developmental timescales.

What makes things 'close enough' is of course relative. In rapid physical actions, feedforward models are used as there isn't enough time for a feedback loop to be useful in coordinating action, even though the thing represented might be at hand (§5.2 Feedback and Feedforward). Distant objects need to be represented as they are too far away for events in the object to be parts of a mental operation as they happen (§4.2 Representations). Sometimes, the border between inside and outside may shift: given a slow process that becomes habitual, something that was external

could become internal (see 6.3.5.1 In a relationship below). Something counts as in the mind, then, to the extent that the information is carries are immediately available to experience given the right kind of attention (see §6.2.1 The Feeling of Things). They must be immediate in the sense that we need go through no further inferences to be able to report on the experiences we are having: 'to engage in qualia talk' (Chrisley & Sloman, 2016).

When it comes to how the two hemispheres are related, the corpus collosum acts to transfer information so the virtual mechanisms embodied in the brain don't need to represent what is happening in the other in order to function. This division of labour is obviously efficient, and continues to operate in cases where subjects have followed a normal developmental trajectory and both hemispheres can access the same information through having access to the same portions of the environment. When they receive different input, in effect, they become different minds, as they no longer share information between processes distributed between them. In cases where humans are born without the connective tissue between the hemispheres, e.g. Kim Peek, then the mechanisms would develop separately. In Peek's case, it seems he did develop language processing function in both hemispheres, and could read two texts at the same time (Brogaard, 2013).

If what matters is being organised in the right kind of way so that information that represents a relevant aspect of the world is immediately available for use by the subject, then anything capable of being organised in such a way will be capable of helping constitute a subject. That is, some kinds of matter (it's an open question whether this is a broad or narrow category) have the property of being susceptible to being so organised. This property of being able to help constitute subjects through being organised in such-and-such a way and involved in ongoing, dynamic processes, explains how mental states can be emergent (irreducible in that they are more than the additive result of the causal properties of the physical parts at any moment - §3.4 Emergence) without violating supervenience (interpreted in a sufficiently broad-minded, liberal manner - §3.3 Supervenience & Realisation).

6.3.5.1 In a relationship

Intentional mental states, as involved in explanations of actions, have an experiential aspect to them, even if that experientiality is not always immediately given in awareness (§6.2 Consciousness). The orthodox understanding of conscious experience is that those experiential properties are intrinsic properties of whatever has them, e.g. a brain (§6.2.1 The Feeling of Things). The account being advocated here maintains that experiential states are not intrinsic but relational, in that they depend on organisation and functionality.

Some indications that the felt properties of experience might, despite appearances, be relational, come from work on sensory substitution (Bach-y Rita, et al., 1969; Bach-y-Rita & Kercel, 2003). In this work, information about the environment that we normally get through the eyes is substituted with information taken though a device (e.g. sonar) and provided to a subject through another modality, e.g. touch. In the original experiments, a sensor was worn on the face and the information about the world was relayed through an array of pins arranged on the abdomen. Subjects report that at first they feel a strange prickling on their abdomen, but that later they learn to interpret this as information about the world. Eventually, they no longer feel pricklings on the abdomen, but report having a sort of visual experience of the world (Poiriera, et al., 2007). This indicates that the way something feels to us is at least partly to do with the way the information is used by the system, that is, its extrinsic, relational properties.

This description of the feeling of mental states being (at least partly) determined by how that information relates us to the world and is generated by our actions brings to mind the idea of active perception: 'Visual perception can now be understood as the activity of exploring the environment in ways mediated by knowledge of the relevant sensorimotor contingencies' (O'Regan & Noë, 2001, p. 943). Noë gives the work on sensory substitution as evidence for the enactive view that perception and action are inextricably bound; that perception is not merely a passive registration of information arriving at the sensory surfaces, but the result of the actions we perform in order navigate the world. If using tactile sensations for actions involved in navigating a space lead to the type of spatial experience usually associated with vision, then how we experience the world depends on the kinds of actions perceptions are used for, rather than being intrinsic to the sensory organs that are stimulated. As Prinz (2006, pp. 5-6) points out, though, the behavioural evidence (learning to negotiate spaces) is not sufficient to say that the experience is visual rather than tactile, and if a person who has been blind from birth reports vision-like experiences, it is difficult to give such claims much credence, as they have no experiences of vision on which to base such a claim.

Other evidence that how we experience the world depends on how the information we get about it is used, comes from perceptual adaptation, where subjects adjust to wearing glasses that turn the world upside-down visually, through a process of acting in that upside world, until eventually the 'upside-down' world seems the 'right' way (Richter, et al., 2002). Phenomena like this may be explained by cross modal expectancy effects, that is, the way we experience something could also depend on information from other sensory modalities, if it is used to help form expectations about what we will experience. For example, the McGurk effect, where visual information affects audition (McGurk & MacDonald, 1976). Likewise, expected motor responses can influence how we perceive,

but this is not to say that the processes involved in motor responses are constitutive of visual experience (Prinz, 2006, pp. 5-6).

What this evidence points out, then, is not that all experience is based on mental processes involved in action, but that feedback over time from our actions in the world constitute expectations that play an important role in forming representations of the world, which are used as the basis of actions, and which partly constitute experience. This is a causal, dynamic process, with multiple mutually influential streams. Motoric action and its possibilities is one of those influences, and one deeply entrenched evolutionarily. However, it may actually be the ability to not have action directly linked to perception and action routines, to think off-line, to visualise, simulate, and plan, that marked a significant moment in the evolution of intelligent creatures like us (Prinz, 2006, p. 11).

It may be the case that we develop the ability to have thoughts about objects in the world without those objects being there to participate in that thought, but the fact is that it wouldn't be the thought it is without in the first instance being constituted by a kind of relationship between mind and world. The internal portions of these experiences of the world are used in feedforward, expectational systems, and in the cognitive processes involved in planning, forming and performing actions, but this does not mean that when we are having an actual experience that only that inner portion is involved in our experience. We are not experiencing that inner state; we are experiencing what it is related to, which involves that inner state. In the case of misperceptions, where the thing we think we are perceiving is not there, we are tokening the inner mental state that is usually tokened when the thing is there, but we are not perceiving the thing itself, rather we are perceiving the world *as if* that thing were there. This misperceiving depends on the cases of actual perceiving, and actual perceptions depend on more than just the tokening of the inner state.

If it is possible that minds can be divided in two and (re)combined given the way information is shared and used, there seems to be no reason why more than two 'proto-subjects' could be so combined. Does this lead to a slippery argumentative slope to a form of panpsychism? We can head off this challenge by putting certain conditions on what makes something the kind of thing that can be a mind. The division of a mind into parts that have the potential to have experience has to 'bottom out,' so we should be able to say something about what kinds of things can be minimal experience producers. If being a subject is not a matter of having a single place where all the intrinsic properties of experience come together, what is it that makes us feel like a single self? In the view being presented here, we are using the notion of a 'virtual machine' to do so (Sloman, 2011a). The mind is composed of multiple virtual machines, which taken separately are simple individual processes, but which together form an experiencer (c.f. Minsky's (1986) Society of Mind).

To restate the claim made at the start of this chapter in light of the above, recasting the relationship between mental states and the physical stuff in which they are instantiated in terms virtual machines has several clear advantages over more traditional functionalist accounts. Virtual Machine Functionalism (VMF) doesn't require the identification of a 'single well-defined state of the system, or a collection of states' (Sloman, 2013) with the relevant mental state. Instead there may be multiple, overlapping (e.g. sharing informational resources) mechanisms that come into existence, form coalitions, and go out of existence. These can utilise neural, bodily and environmental resources, with information flowing in and along multiple channels, rather than following a linear route through a step-wise process. Instead of abstracted mental states realised in discrete physical states, where these physical states have the appropriate causal properties to fulfil their functional role in transforming the preceding state into the consequent state in a linear, input-output process, VMF explains the causal powers of a system by referring to 'a multi (virtual)-machine architecture in which several enduring states with their own causal histories interact causally with each other and with the environment' (Sloman, 1992). Thus, causal explanations refer to states of these machines and how they relate to each other, without assuming that the 'real' causal story is happening at the 'bottom'. The function and operations of virtual machines are implementable physically, but not definable in the language of physics (Sloman, 1992), and in terms of the metaphysical theory here outlined about the relationship between levels of description, requires reference to the temporal causal trajectory of the mechanisms. Causation can go 'upwards, downwards and sideways in virtual machinery' (Sloman, 2013). This fits with the picture of causation outlined above (Chapter 3: Levels of Causal Explanation). Descriptions at the physical level don't 'trump' those at the design level, quite the reverse, since the causal trajectory that explains the functioning of the virtual machine happens at the level of the machine, in terms of the circumstances leading up to the current state. Those circumstances include facts 'around' (at hand, or causally proximal) the machine, whether that be 'below', 'above', or 'about'.

At this stage, we could drop the talk of levels all together; everything is happening at the same level in that everything is physical and is causally efficacious because it is. Talk of new levels coming into existence explains the smell of dualist spirit that puts people off; what emerges over time is coalitions of physical stuff that together exhibit causal properties that it is useful to refer to as wholes, because the wholes have causal properties not explained by the causal properties of their parts when viewed as isolated individuals at a point in time and space. For the moment, however, I will continue to follow the convention of referring to such emergent properties as being at a 'higher level' than physical properties.

Understanding these ‘high-level’ principles allows better understanding of how such systems work for the purpose of explanation, prediction and engineering. Unlike traditional functionalism, which was often interpreted as being only interested in the so-called ‘design-level’ descriptions, with the rest being ‘mere’ implementation detail, understanding how virtual machines work, what they are built to do, and how evolution, development and society formed them, is a piece of interdisciplinary, empirical and conceptual work with clearly defined objectives. The ‘design level’ is not a separate realm where there is a blue print waiting to be materially realised. This is to confuse the way we design and build with the way nature evolves. Nature’s process is incremental changes with no direction. Traditional functionalism would take a problem domain, design a solution, then engineer it. So we design a navigation machine by using senses to build a model of the environment, use the model to plan a route through it, then use the plan to set in motion a series of motor actions. Then a ‘full-on’ functionalist may claim that any system that behaves just like this navigation system is an implementation of that design. However, it may well turn out that that’s not how nature has solved the problem, since its solution was built randomly and incrementally on top of pre-existing systems. So, understanding how nature has built navigation systems may lead to novel empirical predictions about the behaviour of such systems, particularly patterns of failure, and could help us build better navigation systems ourselves, since our solutions may be too computationally expensive. Thus, virtual-machine functionalism, unlike the traditional type, is a natural framework for the interdisciplinary endeavour that is concocting a scientific understanding of mind.

Chapter 7: Natural Minds

7.1 Embodied Virtual Machinery

Our subjecthood is a datum to be explained rather than an unshakeable foundation on which to build a system of knowledge. It has been my aim to provide at least the beginnings of such an explanation, one that grounds our understanding of ourselves as unified subjects acting for conscious reasons, and brings closer together the normative, reason-giving style of explanation and the causal style of explanation.

What makes me 'selfish'? Firstly, it is the automaticity with which the various parts of my mind yield their contents to the whole. This makes the parts work as a 'systematic whole' with no noticeable delay caused by obvious processing. Secondly, the 'lower' processes make their contribution without being available to introspection directly; we are only aware of the 'finished product.' This does not imply scepticism regarding the 'higher' contents of our mind that we use belief-talk of, but rather a critically realistic revisionism, one that helps rid ourselves of philosophical baggage that taints the very introspections that we use as data to justify our sense of self.

Without the close integration of the various parts that make up our minds, we would not function as organisms, or survive as a species. This selfhood is not an illusion; it is a necessary feature of creatures with complex brains because they have complex lives: our decision-making mechanisms have evolved to enable us to act on salient inputs. The ontological level at which to seek explanations of actions that result from these mechanisms is that of those aspects of the world they have evolved and developed to operate on. A mechanism that enables us to learn and apply the norms of the social groups we operate in will be best described at the normative level. Such a machine may be built in many ways, where what these ways share is that they instantiate the same 'machine.' There will therefore be no way to reduce machine descriptions to physical ones, even though the physical description necessitates the machine description in each case (maintaining supervenience). Since functions like recognising threats lead to actions like taking evasive action, there will be downwards causation (as well as upward, and horizontal). Kim's (1992) arguments against downwards causation (§2.1.1 The 'Special' Debate) will not work against multiply instantiable machines, as opposed to multiply realisable states, because 'virtual machine supervenience' requires the existence of a 'network of causally interacting, *enduring*, components' (Sloman, A., pers. corr., 23/8/2011 – my emphasis). That is, machines are necessarily diachronic entities and Kim's arguments only tell against synchronically described states (Chapter 2: Physicalist Reductionism).

Virtual Machine Functionalism is more demanding than simple functionalism, in that it is not a 'black box' theory: it is not enough just to say 'whatever fulfils this functional role,' when specifying what comes between input and output; a machine design is specified, the causal properties of the whole being determined by how the parts work together. Actions are explained by referring to these machines, and by how those machines were constructed, which includes what they are made of, and the processes that formed them. The fact that a machine is a dynamic whole means that being in time is an essential feature; unlike the laws of physics, the generalisations we form to understand the causal consequences of such machines are not time-symmetrical. In general terms, actions are brought about by 'need sensors' (e.g. for sustenance), and 'fact sensors' (e.g. presence of food) (Sloman, et al., 2003, p. 10). Needs that are stored internally and represented to the organism without having to be directly stimulated by the environment may be called 'goals' (Sloman, et al., 2003, p. 12). The rest of this section will look back at the metaphysical questions discussed earlier and show how this view may answer some of the questions raised.

7.1.1 Causation & Causal closure

The mind is made of multiple concurrently running virtual machines, each with causal relations to other virtual machines in the mind and portions of the world, as well as phylogenetic and ontogenetic causal histories. These causal relations are necessarily diachronic in that to account for and explain the output, including how that output fits into a larger pattern of behaviour, it is necessary to cite events in the past relative to that output. The statements of causal connection will not be strict and hedged with *ceteris paribus* clauses. The causal properties of these virtual machines will be probabilistic and dispositional by nature, and the resulting causal event highly context dependent. Each machine is part of the causal profile of a context, and when we refer to its causal contribution, we are not citing a universal regularity that is being prevented from expressing itself, but rather pointing out the fact that that mechanism tends to lead to this result in those contexts. This accords with a 'difference-making' analysis of causation (§3.1.3 Making a difference).

This view of the causal nature of the mind undermines the version of causal closure most commonly used in arguments for reducing the mental to the physical, because it is temporally and spatially extended. Machines, even virtual ones, take up space, and work in a sequence of parts affecting each other. Take as an example the way attentional mechanisms function in presenting the world to us in certain ways. Changes in perception are caused by changes in attention, which is affected by, amongst other things, knowledge (O'Regan & Noë, 2001, p. 970; Sloman, 2011b). We learn to notice salient forms in the world around us, which are picked out from the flux and presented to our

conscious awareness. This may be because of basic drives, e.g. for sustenance or sex, drawing our attention to relevant bits of the world. Even in such instinctual cases, where we as agents may have limited control over what our eyes are drawn to (although we might have varying degrees of impulse control thereafter) there may be associative learning involved in seeing objects as potential 'satisficers.' In other cases, it is clearer that attention is driven by higher-level knowledge systems, for example when searching for a particular word in a page of text.

In these cases of 'top-down' control, from the contents of conceptual beliefs to perception, these are instances of downward causation (§2.1.1 The 'Special' Debate). Why did my eyes move that way? The causal story involves a mechanism which requires time to function (and to evolve and develop), which violates a strict, synchronic, local causal closure principle (§2.2.1 Causal Closure). Of course, there is a narrow type of explanation that involves such causal processes as wavelengths of light, retinas, nerves and neurons. These types of explanations are good for some kinds of questions, but will not satisfy all the 'why' questions we can legitimately ask about actions. If society morally judges someone for a failure of impulse control, a microcausal story about the molecules in their motor cortex will not be sufficient; we will need to refer to larger scale dynamics at the person level: we blame a person, not a neural circuit (although a neural circuit gone wrong, i.e. not functioning normally, may be reason not to hold a person responsible for their action).

7.1.2 Natural Kinds

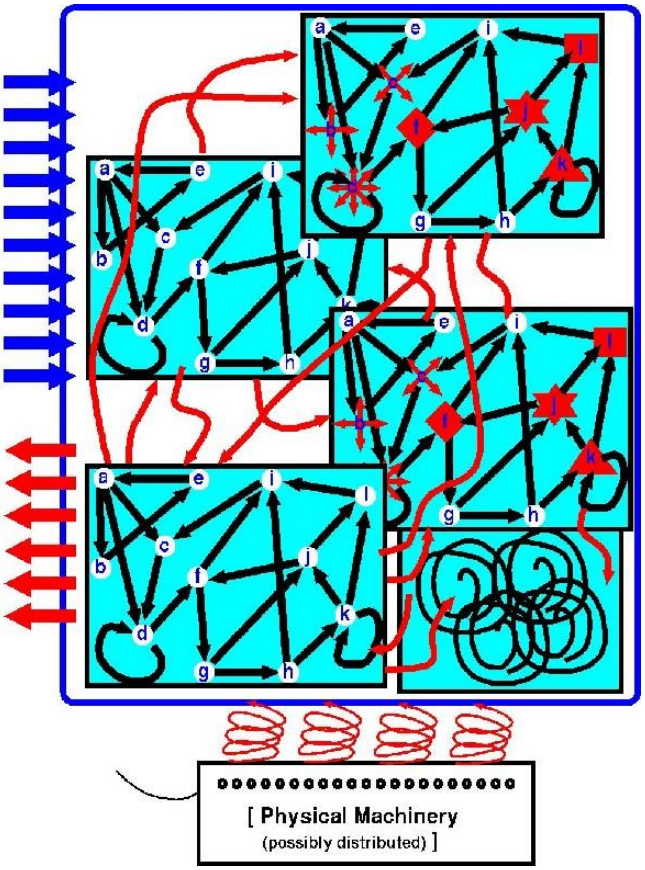
When it comes to the kinds of things referred to in explanations of intentional action, microphysical kinds will also not suffice. The causal story involving those kinds of things may be able to explain why a bunch of molecules (which happen to be in your hand) moved in a certain way (which happened to be upwards), but it wouldn't be able to explain that the reason for that movement was that you were reaching up to pick the fruit off the branch. Nor would the microphysical focus allow the formation of generalisations that could support predictions of future similar actions in similar situations, i.e. feeling peckish and seeing some apples ripe for the picking. Events described this way would not be 'visible' under that sort of explanatory 'lens.'

Science gives us understanding to the extent that it explains observations, and as such its statements need to say something about the observed event in such a way that that understanding is transferable to other events of sufficient similarity. It doesn't really have a place for singular causal statements, as these are too particular, never-to-be-repeated. In microphysical terms, every 'macro' situation is that particular in the real world, only in the isolation of the laboratory can we create enough similarity between different occurrences so as to form generalisations, so, in order to form

generalisations useful in the real world, we need to be able to talk about an ontological level that includes objects like organisms with mental states in a world of objects. The terms used in scientific statements need to pick out things that are shared by various contexts, and so need to be projectable, that is, they need to be kind terms in the sense described above (Chapter 1: Natural Kinds).

Given VMF, we can say more about the kinds of processes in operation than the abstract causal roles of traditional functionalism. In fact, we can say enough about them to allow empirical sciences to get hold of them in a way that wasn't possible in the purely formal language of old-fashioned functionalism. It is here that the topographical view of natural kinds comes in useful. The formal language of computationalist functionalism is so refined, the boundaries of types as picked out by symbols with purely syntactic properties so 'clean,' that it is of only limited use when dealing with the real, messy world of stuff and things. In fundamental physics, the question of whether something is or is not a particular kind of particle may be clear, and given enough computing power the interactions of these may be predictable to arbitrary degrees of complexity. But, these outcomes will be unique in their particularity. From a higher level, we see patterns that are bigger and 'messier,' heaps of matter bound together by causal histories, more or less loosely, about which we can make useful generalisations because of the dynamics they share. VMF is not fully abstracted, as it can point to particular kinds of real-world, embodied processes that play the necessary roles; it is only abstracted enough to allow that the different patterns of physical causes that play the role of, say, indicating that the apple is ripe for picking, are 'lumpable' together as beliefs with the content that the apple is ripe for picking, and in particular circumstance, e.g. when linked with another state that in our respective cases is the feeling of hunger, will lead reliably to the action of picking an apple.

A virtual machine is still abstracted, but with the real world in mind, as it were. As an illustration, take the following diagram, which shows how such a machine is a complex mechanism for dealing with complex input and output, and, moreover, one that is physically realised in a way that is flexibly distributed. Each of the component mechanisms, represented by letters, should be simple enough to specify its workings, with the arrows representing information flow between them. The short input and output arrows on the left represent some potentially complex and distributed, while the coiled arrows at the bottom represent the physical machinery that realises the virtual machines and constrains the machines that may be realised. Each part is simple and physically realisable, but taken together, if assembled in 'the right way,' should be sufficient for the kind of mind I have been describing here.



A Multi-component Virtual Machine, with components un-synchronised, new components coming into or going out of existence from time to time, and some of the components discrete, others continuously variable, all of them running on a Physical Machine, which may be composed of multiple networked physical machines, like the internet…. [and] the components of a VM may be parts of causal loops passing through the environment, for example during control of physical actions. (Sloman, 2013)

This machine description is abstract in that there will not be a definite physical description that all machines of the same type follow, given that the same function could, in theory at least, be realised in different ways. It's grounded in the real world in that its design is constrained by the materials at hand and the time needed for the processes to construct them. In the case of human mental processes, that means genes and other reliable developmental resources of the physical and cultural world, and the materials of the brain and the body over evolutionary and developmental timeframes.

Our question is whether there is room in the world of natural causes for things like those referred to in standard belief/desire type explanations of intentional action. I believe that the foregoing arguments point towards an affirmative answer. States that we can call beliefs and desires (or belief-like and desire-like) are physical states of a virtual machine that play the role of carrying useable information about parts of the world, and motivate certain forms of action in the world, using the information carried by the belief-like states. The process by which those states come to be what they are is one of multiple feedback mechanisms over phylogenetic and ontogenetic time. For example, a state comes to represent the presence of a predator because it is reliably caused by said presence, and in the past has permitted the organism and its predecessors to escape predation, in conjunction with a motive state that upon detection of a predator leads to behaviour that promotes

escape. Clearly, some such states will be more or less innate reflexes, e.g. a mouse's fear of snakes, while others will be learnt, e.g. a human fear of guns. These states are causal by virtue of being parts of a physically embodied machine, but their causal powers are powers they have in virtue of being belief-like and desire-like states of the virtual machine. They are natural kinds in that we can form true, useful generalisations using them; we can predict the behaviour of creatures that are classifiable as rational because their action is based on inner states that are there for a reason. This reason, which is that we can tell a good evolutionary story as to why they are there, is what makes them not merely things about which we can take the intentional stance towards. They are things that we can take the intentional stance towards because they have inner states that have evolved to perform the function of representing the world in a way that is useful given the needs of the organism.

There is a possible objection to the VMF view of mental states as natural kinds, which is that it is through direct, first-person acquaintance with mental states that we pick out mental states. The behavioural, observational, third-person regularities are both causally and inferentially consequent, rather than being reference fixers in themselves. The problem is that if we are directly acquainted with mental states in a way that renders their essential properties fully transparent to us, then it cannot be the case that certain opaque properties of mental states, like physical or functional ones, are essential rather than contingent. We can respond by questioning the notion of acquaintance, as it assumes this transparent access to the essential properties of some objects. Papineau does this by assuming that all atomic concepts 'are related to reality by facts external to our a priori grasp, such as causal or historical facts' (Papineau, 2006, p. 102), relying on a Kripkean causal theory of reference by saying that this is 'direct' reference.

While I agree with the aim of seeing through 'transparency,' I think different meanings of 'direct' may be being conflated in saying 'both pain and C-fibres might similarly refer directly to the same entity, yet this not be transparent to someone who possesses these concepts' (Papineau, 2006, p. 103). This assumes a kind of parallel relation between these terms and their objects. But there is a difference. Whereas we can 'baptise' objects ostensively with a name, leaving the empirical work of discovering just what makes things of that kind what they are to later, it is not so clear that we can do this with terms like 'pain,' as this assumes a shared space where ostensive definition can happen. Words are public, whereas feelings are not. For a Kripkean account to work, we need to be able to say 'let's call things of *that* type such-and-such,' and for this to be understandable by other users of one's language.

Rather, I think the distinction between kinds of kinds like chemical kinds and mental kinds is this: we can discover necessary facts about the physical essences of chemical kinds we initially pick out by naming a particular sort of stuff we find around us in the world. But with mental kinds, we learn them by observing the behaviour of others in particular contexts, identifying what happens in ourselves in similar contexts, and then discovering what realises these in this world contingently. Given all the facts about salt, and all the facts about sodium chloride, the entailment is necessary in both directions, whereas the same is not true of mental kinds and their realisers. We never discover that 'pain' is identical to C-fibres firing, we discover that pain happens to be realised by certain physiological processes in us, but that doesn't necessitate anything about how it may be realised in other creatures with different histories. Even if it is insisted that the pain=C-fibres firing identity is token rather than type, it is still too direct an identity, as it is not C-fibres firing that makes it pain, but a broader fact about the mechanism that the C-fibres play a role in.

Therefore, we may accept the use of the term 'water,' through a process of refinement that includes the conventional acceptance that science has authority over fixing what counts as water, as referring to stuff that shares the physical properties of being composed of H_2O molecules. Given that the lawlike generalisations we make about water are necessitated by its chemical composition, and exhaust all the generalisations we may care to form, XYZ can be ruled out as a possibility, since there will be some laws that don't hold of both substances, otherwise they would be the same stuff. For example, the law that at very high temperatures, when the water molecules begin to dissociate (about 2000°C), there will be hydrogen and oxygen gases present. The result is that the XYZ world is not a possible world after all, because water is not a functional kind. However, mental kinds are functional, and the worlds where the same mental kinds can nevertheless not share the same physical essences are possible. They are what they are because of what they have been selected for, and indeed they have the properties they were selected for because of their physical properties, but it is not those physical properties they were selected for, and other physical properties could play the same role.

The same point can be made about representational states. A mental state that represents something in the world is more than a merely causally covarying brain state, it is a causally covarying brain state that has been selected because it covaries with something in the environment that the organism finds it useful to have information about. It was selected because of its having physical properties that allow it to play a particular role. As such, it is the kind of thing we can make useful generalisations about, but because many other physical states could have played the same role,

there is no necessary identity between narrow micro-physical instantiation and the mental state that is instantiated.

To be called a law, according to Goodman (1955), a generalisation must be 'projectable,' that is, 'confirmable by its positive instances and supportive of subjunctive and counterfactual claims' (McGinn, 1978, p. 204). But the physical essences of mental states are not so projectable. There are unlikely to be any useful psychophysical laws because the variability of the realisation of mental states by physical states precludes an explanation of the coincidence of these states. But we don't need psychophysical laws for mental states to be kinds if we reject the requirement for physical real essences. In each case, the realisation base can give an explanation of how the mental state achieves the causal properties that make it the kind of mental state that it is, but that state's membership in the broader functional category it belongs to, and which justifies the projections we can make on the basis of this categorisation, is based on its function within the system, which includes organism, mechanism, and environment. It is in providing this higher-level similarity space in which to 'lump' otherwise distinct physical types together that the topographical approach is an advantage over other characterisations of natural kindhood, characterisations that lead to anti-realism about natural kinds because of their reliance on strict physical essences and clean boundaries.

The relationship between physical instantiation and mental state is a broad supervenience conditional, in that we can say that if a certain physical state is instantiated (not a narrow neural correlate, but also including other parts of the body and world that are part of the functioning of that state), then a token mental state of that functional kind will necessarily be instantiated (§3.4 Emergence). We don't need strict psychophysical laws to legitimate a realist stance regarding mental kinds; the existence of true generalisations at the functional level is sufficient, the truth of these generalisations being underwritten by a theory of causation that can accept causal statements at this level (§3.1 Causation). These causal statements are diachronic, rather than synchronic as in 'pure' physical causation, are not affected fatally by overdetermination, as they are different, equally valid, descriptions of the same cause (different in that each satisfies a distinct explanatory interest), and are respectably physical in that they do not violate the transitive causal closure of the physical world (§2.2.1 Causal Closure).

7.1.3 Physicalism

Being physically realised, the kinds of systems in question, ones that contain causally efficacious intentional states, will always have a complete physical description, and such descriptions may be

used for some kinds of predictions, depending on our interests and epistemic and computational limitations. Such descriptions will miss larger-scale patterns because they lack the vocabulary to pick them out, and will be limited because of the inherent indeterminism at that level. Thus, higher-level descriptions that capture larger flows of causation at work, may, in certain situations, have explanatory priority. This is metaphorically like the relationship between currents and liquid: there is nothing more than the matter in the liquid, and the motion of each molecule in isolation is a function of its previous state and the effects of its neighbouring molecules. But, taking a wider view, we can see patterns of currents caused by the broader context, and we can see similarities between types of context, i.e. those that tend to lead to whirlpools (which is the kind of embodied knowledge an expert canoeist comes to have). Looked at in terms of difference making, it is these contextual factors that cause certain currents to emerge, and it is those currents that push the molecules of water around. The properties of the molecules of course play an important role in constraining the kinds of dynamics permissible, but for the purposes of explaining fluid dynamics of water, those properties become background conditions rather than causal contributions.

This version of physicalism is not particularly esoteric. Physicalism is about constitution rather than causation, and 'straightforward physicalism' can be characterised as using the term 'physical' broadly, not just to refer to kinds studied in physics departments, so 'it should be understood as including physically realized role states along with strictly physical terms' (Papineau, 2006, note 1).

Sloman, when describing how he sees VMF as regards physicalism, makes a stronger claim:

> … a consideration of examples of virtual machine functionalism where the virtual machinery includes several concurrently (but not necessarily synchronously) active sub-systems with their own causal relationships, both within the VM and also across its boundaries (e.g. to internal physical memory, to physical interfaces and even to things referred to in the environment) would show that there is no version of physicalism as normally defined that survives, even though all the virtual machinery is ultimately implemented in physical processes. Part of the reason for this is that patterns that can exist in physical structures and processes need not all be definable in the language of physics. For example, what makes something an expression of the English sentence "Today is Fred's birthday" depends not only on how current users of English read text but also who Fred is, which calendar is in use and other complex social facts. What makes the state of part of a virtual machine a case of someone wondering what Fred thinks of him is even more remote from being translatable into a physical description….

I would agree with everything except that I think I have described a fairly straightforward version of physicalism that does survive.

7.1.4 Emergence

The view advocated here is that the mind emerges over time through the feedback dynamics between organisms and environment found in biological and cultural evolution, and individual development. Part of this individual development involves conscious reflection and the freely willed actions that result. It is this, partly, that drives the emergence of novel behaviours which are then available to be copied by others in our social groups if they are judged to be worthwhile. These behaviours are caused by emergent belief-like and desire-like states, and it is by reference to these states that science can sometimes give true explanations of behaviour, and furthermore, make good predictions.

This emergence is robust in that the states in question exist with sufficient stability in terms of time and space, from our human perspective, for us to use them for meaningful generalisations; if we were very different creatures in terms of the timespans we operate in, or the kinds of things that are significant for us, then we would arrive at a different description of the topography of the world of objects around us. There is not one unique such list of objects or one privileged perspective. Furthermore, this sort of emergence is not problematic for physicalists because it is not claiming that the novel objects that emerge are eternal kinds that need to be added to the pantheon of fundamental things that make up the universe. They exist within a four-dimensional bubble in the transitive causal closure of the physical world. Such systems are partly definable by the property they have of gathering resources to stave off the entropy of the second law of thermodynamics, but nothing can hold that off forever.

7.2 Mental Causation

7.2.1 United We Stand

'Why did I do that?' is a seemingly simple question. If we act intentionally, we give reasons we consider to be sufficient to rationalise the action. Our interest is in the cases, if any, such explanations are true, rather than rationalisations of non-intentional behaviour, or merely metaphorical ways of referring to physical processes. What would make such explanations true is having veridical experiences of decision making processes that involve believing certain things (facts and norms) and desiring particular outcomes. My goal has been to dispel the apparent tension between this everyday conception of ourselves, as intentional beings who are the causes of our own actions, and the reductionist tendency that assumes that all the real causation is describable, in principle, at the level of fundamental physical stuff.

How can we reconcile the explanation of my action that involves the causal story of neurotransmitters, etc., with the explanation involving conscious reasons? There is an apparent 'gap' between these explanations, in that one cannot be derived from the other, and a tension, in that they appear to compete for consideration as the genuine cause because they cite, as *explanans*, states that seem to belong to different ontological realms. To close this explanatory gap, I have tried to reduce the epistemological distance between felt experience and the physical world, and to render unproblematic the view that there can be different levels of genuine explanations within a physicalistic ontology.

Organisms like us track the state of the environment in some detail and with some robustness, moreover we monitor our perceptual states and inner states, and we act on this information in a more or less coherent fashion using our limited supply of energy and mental resources. We are not merely reactive, but flexible and adaptable, and unlike simple computers we do not follow algorithms, even if much of our behaviour may be describable algorithmically. We process information, but in a way that doesn't lead to automatic behaviours triggered by events, rather information is processed and an action is initiated using stored energy (Sloman & Chrisley, 2003), which we harvest from the environment through eating and breathing.

Purely reactive systems following algorithmic processes may in principle be able to produce all the behaviours that deliberative ones can, but that would mean storing all possible scenarios and reactions to them, which is not practically possible or evolutionarily feasible. Given the variation and complexity of human environments it makes sense to have a mechanism that can solve novel problems based on similar previous experiences, or pure creativity. This requires a reflexive 'meta-management' that monitors the system and prevents it getting stuck in indecision or otherwise 'crashing.' This self-monitoring process may involve the need to categorise inner processes, and be the basis of the kind of self-awareness we call consciousness (Sloman & Chrisley, 2003, pp. 155-9).

These layers of cognitive processes can be seen as emerging over evolutionary time, building upon what has been already achieved in an arms race where the winner is the organism with the greatest level of freedom to adapt. We have achieved a level of evolution where we can reason abstractly using concepts we learn and then work with and on. But this is still a work in progress; we are limited beings, and our concepts are 'tethered' to our embodiment. I take this to be what Hurley (2003) meant by describing our space of reasons as locally coherent systems of concepts bound by context rather than fully generalizable. Some of our conceptual abilities may approach the fully general, and it is to attempt to achieve those heights that we engage in scientific philosophy, even if this must ultimately remain an ideal given our limitations.

However, when we speak of the fully general, it is important to remember that we are not referring to approaching the drawing of the one true map of a territory. The topography of the conceptual systems we create will always be relative to us, and our interests at that time, including the level of detail we need or can usefully use. But given both those constraints, there is a standard of correctness, which is fully constrained by the actual structure of the world. When we use our maps, we can be said to be wrong in a way that a moth circling a light cannot. The causal structure of the environment has changed, meaning that its design is no longer fit for purpose, but it has no beliefs about the source of the light to which we can apply normativity. Our deliberate actions, on the other hand, are based on models built using feedback, and these models can be said to be wrong because they are constructed using concepts, and the model may not accurately map onto the world.

A possible objection to this causal interpretation of intentional explanation is that we may be eliding the distinction between a causal interpretation of folk psychology, where folk psychology is taken to be a largely true theory about the overall architecture of the human mind (*a la* Fodor), and the interpretationist interpretation, where folk psychology is seen as (merely) a practical tool for prediction and coordination (*a la* Dennett). Sterelney (2003) & Godfrey-Smith (2003) claim we are faced with a dilemma: if we accept the causal interpretation it is hard to understand how such an architecture could arise by gradual mutation; if we take the interpretationist route, then folk psychology doesn't cut our architecture at the joints.

On the first horn, Sterelny gives strict condition for something to be said to be acting for reasons of its own: it must track its environment richly and robustly, this tracking should be largely decoupled from specific responses, and its motivational states have to be informationally sensitive rather than being based on the relative strength of internal drives (Sterelney, 2003, p. 263). However, as should be clear from the preceding, this is much too strong a claim; it's an idealised abstraction with as much real-world application as classical economic theory. To argue from this abstracted ideal of rationality to the impossibility of evolving gradually to this point is to make the same mistake as creationists in assuming a perfection of design that is not evident. If the formalism of the logical systems we use to express our best reasoning were actually an accurate way of describing the architecture of the mind, then we would find formal logic as simple as riding a bike, or walking down the street. If instead we accept a more local, less abstracted holism when it comes to the reasons we give for action, then it is not difficult to see how these abilities could evolve gradually. Moreover, this middle way allows us to pass between the horns of the dilemma, as interpretations

are acceptable as long as, once we 'fix' the points of reference, there is a way of judging accuracy against reality.

It may be that there is no real conflict between the causal and interpretationist views. If, as Sterelney (2003) allows when sketching a possible way to take up his challenge, one of the important and distinctive facts about us is that we are subject to cultural as well as biological evolution, then interpretations can have a causal role. We are designed to be able to develop in a social niche, where much of our learning is based on interpreting the actions of others and internalising what we learn, making it our own: 'The connections to natural and social environments, as well as the internal wiring, are essential to the causal powers of minds themselves. This is true at both the personal and subpersonal levels, at both the level of content and the level of vehicles of content' (Hurley, 2003, p. 275).

In arguing for taking the mental states referred to in intentional explanations as natural kinds, I am claiming that the general framework of intentional explanations that include belief-like and desire-like states ('how things are states and what I want states' (Hurley, 2003, p. 271)) provides a genuine causal model for intentional action, one that can ground scientific explanations and predictions, and is open to scientific exploration and refinement. Our ability to drive our own actions based on our own reasons has been woven through the intertwining feedback dynamics over phylogenetic, sociogenetic and ontogentic time. The intentional level emerges through the development of increasing flexibility, including the development of tool use, until the symbolic 'explosion' and the resulting ratcheting up of cultural evolution. These feedback systems form 'dynamic singularities' which persist and reproduce, and provide non-reductive explanations of the actions that result.

To the question of what mental causation adds to the world if mental events are token-identical with physical events, we can answer as follows: without the 'binding' of the disparate physical events by coming under a statement of mental causation, the physical events at the end of the causal chains would be seen as mere coincidences. This is downwards causation in the sense that we can infer lower-level, physical facts from higher-level, mental ones, and we can give explanations for physical events in terms of preceding mental events without fear of making category errors. These mental causes are not external to the physical causal order, just that they are invisible from the perspective of the causal properties of the composing particles. If one were to somehow 'bracket' the physical parts that constitute a mental state without referring to that state, and to trace the causal trajectory forwards or backwards, the result would be a complex case of singular causation (Lowe, 2000, p. 584). It wouldn't be very useful to do this, as explanations like this would

'leak,' in that the trajectory of some of the parts would not be part of the constituent matter of the subsequent or consequent mental state (if we were to take a step back).

Being realist about intentional action is to take it to be a kind of causal explanation open to counterfactual analysis (Heyes & Dickinson, 1990, p. 106). I believe that I have developed a realistic account of intentional action, based on three main premises. These are: a theory of causation based on the causal tendencies of contextualised objects in the world; a subtler understanding of the principle of causal closure that retains the essence of physicalism without begging the question against supervenient causes; and an account of natural kinds that respects both the chaotic, partially undifferentiated nature of the physical world and the role we play in bringing order to our experiences of it. This places the science of intentional action firmly within the interdisciplinary field of cognitive scientific pursuits.

7.2.2 Freedom!

I will end with some comments on the implications of the foregoing for the idea of free will. The theories presented here, I argue, support the common-sense idea that we are creatures whose deliberations lead to actions. When a person is trying to get to the top of a mountain, there are states in her which are motives and motivational, as well as knowledge and skills. These explain her actions both by playing a role in causing them and in rationalising them. Moreover, those states are there because they have that function. And to the extent that those reasons can be said to be her own, then the action that results is her choice, i.e. she is there of her own free will. In the case of the water droplet 'trying' to get to the bottom of the window, there are no such states the presence of which is explainable by their having the function to achieve such goals. Therefore, the use of intentional descriptions of drops of water is purely metaphorical.

Of course, it is important to realise that we are not always acting freely when we think we are; we sometimes confabulate or mislead ourselves, or are manipulated and misled by others. Equally important, though, is that this does not mean that there are no cases in which our sense of agency is veridical. In fact, using cognitive science to better understand in what ways we can be wrong about whether we are in control of our actions is a route to greater freedom, as this understanding can feed back into future decision making processes.

An example is the bystander effect (Darley & Latané, 1968), where a subject comes across a situation in which someone is lying on the floor, apparently unconscious, but does not intervene, as they take their cue from others, who are not doing anything. The more bystanders there are, the stronger the effect. Subjects will rationalise their behaviour to maintain their view of themselves as autonomous agents. If so, the reasons the subject might give for (in)action are not in fact the causes

of that behaviour, so said subjects cannot be said to be acting freely. However, knowledge that one is susceptible to the bystander effect can influence how likely one is to fall prey to it in the future, as this knowledge will feed back into future decision making. This might only happen given enough time for slow, deliberative thought processes influence behaviour, rather than fast, automatic processes. Over time, however, given enough opportunities, these thought processes might become habitual.

The issue of determinism is often seen as problematic for a physicalist theory of free will, as it seems to prevent actors from choosing between possible futures. However determinism is not really relevant to the question of whether we can be the source of our own actions in a genuine way. Being able to affect one's future is what counts, and that means being the cause of one's actions. The answer to the question, 'Can there be self-generated improvement?' has 'nothing to do with determinism, and everything to do with design' (Dennett, 2003, pp. 91-92). That is, the question is really whether we can learn from experience and consciously apply that learning in our deliberations, the results of which play a role in causing our actions. It makes evolutionary sense to design learners, as they will eventually outperform non-learners. Freedom emerges as a result of evolutionary and developmental processes, and is also constrained by the accumulations of those processes. We are 'designed' to acquire dispositions, but we do have some ability to decide which dispositions to have. We are creatures of habit, but we can be held responsible (sometimes) for the habits we have.

An interesting example that shows mental causation to be real, and to be necessary but not sufficient for the action that results to be free, is behaviour that results from a placebo intervention. If a subject acts because of a false belief about a procedure carried out by someone pretending to be a doctor, then the action cannot be said to be free because it fails to satisfy externally evaluable norms of rational belief formation; they have been manipulated. This means that being free depends not only on internal facts about mental causes, but also external facts about what counts as a rational belief given the information a subject could reasonably be expected to have. This makes the question of what counts as free relative to a socio-historical context, which some might see as a problem, but which I won't develop here. In the case of such actions, though, at least we can say that the subject's belief was causally efficacious, i.e. non-epiphenomenal.

Of course, the beliefs that cause the placebo effect are physically realised, but, given the arguments above, this should not be seen as excluding the psychological explanation at the subject level. It is a hangover from dualistic thinking that is responsible for the idea that a physical (i.e. neurological) explanation 'trumps' a psychological one. This is clear in much of the scientific discussion around

experiments on free will, e.g. in Libet's (1985) studies and its successors, in which subjects decide when to perform a simple action, noting when the decision to act was made (Libet, 1999). Meanwhile, data is gathered from EEG, showing that there is a build-up of activity before the formation of the conscious will to act. This preparatory activity in the brain is interpreted by some as showing that we do not have free will, as the cause of the action is neurological, rather than the will to act. Criticisms that such simple behaviours as a flick of the wrist cannot be used as a model of free will in general, or that subjective measures of the timings of events are unreliable, have led to further refinements of the technique. In one, subjects had to decide between pressing a button on the left or right, and noted which letter was being displayed on the screen at the time of deciding (Soon, et al., 2008). The experimenters used fMRI scans to see how far in advance the decision could be reliably predicted, and claimed to be able to do so up to five seconds before subjects became aware of their decision. Furthermore, they suggested that the brain starts preparing for actions like these up to 10 seconds before.

The conclusion drawn, that conscious willing is epiphenomenal as the true cause is neurological, relies on an implicit dualism. But, given physicalism, these arguments lose their force: the mental states involved in causing the actions in question will be physically realised, so in order to show that the actions are not freely willed, it has to be shown that the neurological activity is not these mental states, but other ones. But with 'low stakes' actions like pressing a button with no real consequences, and no extended period in which to develop habits of behaviour that satisfy needs, it is unlikely that we will spot the kinds of neural activity involved in actions where the exercise of free will is important, based on 'hard won' beliefs, like the decision to vote in a particular way. There may be some who haven't made up their mind when they enter the voting booth, and whose decision is rather like a random flexing of the wrist, but we can ignore them for present purposes (or more generally). Some people make a carefully considered choice, reading up on the issues, looking at the options, subjecting their own political prejudices to clear minded scrutiny. In such cases, I think we would be more likely to observe physical traces of these considerations playing a causal role in the action of placing a cross in a box. The fact that this preparatory physical activity may be evident long before the immediate causes of the behaviour is not surprising given that these intentions come from a long process developing personal preferences over time.

Such considerations are taken into account by other scientists in in this field, e.g. (Brass & Haggard, 2010; Dayan, 2008), who accept that there may be multiple mechanisms of decision making, an important one being where the outcomes of past actions are evaluated, and fed back into the next time a decision is made in similar circumstances. Of course, this involves physical traces of that

evaluation in the mechanism (e.g. in a physically realised virtual machine), but this is no reason to say that the resultant action is not free. Decisions that consider the results of past actions are rational, and therefore, free. Random, spontaneous actions of no consequence are not free, they are just random.

This position could be called a version of mechanism based reasons responsiveness (Fischer & Ravizza, 1998), which claims that the important causal role in the sequence of events that leads to an action is a feature of the agent (the mechanism) that is responsive to reasons. In other words, for an agent to be free, there must be in the agent a mechanism that is sensitive to reasons in such a way that if different reasons were brought to bear, its response would be different, and would differ in a way that that course of action accords with constraints of normativity. This account of the nature of these mechanisms supports the externalist view of reasons, where it matters how these reasons come about. This avoids problems raised by manipulation cases, where reasons enter the mechanism 'artificially.' For example, imagine a 'swamp-reason,' where some freak physical event brings about a mental state which happens to be exactly like a good reason for acting. The resulting action would not be genuinely free on this account, because it didn't come about 'in the right way.'

The mechanism's sensitivity to reasons also needs to be calibrated in the right way. A mechanism that is too responsive to the 'right' reasons would mean that knowingly doing moral wrong would not count as free, and the subject wouldn't count as morally responsible. On the other hand, if it is not responsive enough, an insane, irrational agent could count as morally responsible. Fischer and Ravizza (1998) aim to show that appropriate mechanisms are ones that respond to 'reasons that hang together rationally as a class and fit a coherent or sane pattern' (Fischer & Ravizza, 1998, p. 62). Such a mechanism would not need to be infallible, so long as it is responsive enough to reasons. On the question of how sensitive is sensitive enough, common-sense vagueness is sufficient; if we have to argue over the borderline cases, this seems like how moral debates do, and should, happen. A judgement that I shouldn't have done what I did, and the guilt that (perhaps) accompanies that judgement, is based on the fact that I recognise there weren't good enough reasons for acting as I did, and that I should have acted on the basis of the good reasons I had for acting otherwise, or could have had given sufficient thought.

The occurrence, within a subject, of causally efficacious mental states that match information about states of affairs with motivations to bring about other states of affairs, together with knowledge of what needs doing to achieve these ends, provide sufficient conditions for actions that can then be judged to be free or not depending on wider considerations, including standards of rationality. We act 'in the light of' reasons, which are the contents of our beliefs and desires, but it is w*hat* is

believed that is grounds for justifying action. A reason for action is a state of affairs, not just a state of the agent. In the case of a man's paranoid belief he is being chased by aliens, that belief is a reason to seek professional help; if he hides in the bushes, he is not acting for a reason that would make his action free.

Conclusion

The aim of this thesis has been to present an empirically informed view of the metaphysical issue of mental causation, and a philosophically informed view on the status of cognitive scientific general statements. This has been achieved through showing that cognitive kind terms are natural, according to a topographical view of natural kindhood. As scientists, we can say useful things using cognitive kind terms, and these statements will not be eliminated through being replaced by statements referring only to the kinds of things the physical sciences talk about. That is not to say that our understanding of mental kinds, and of ourselves, will not be refined by physical-level investigations, or indeed by discoveries at the level of facts about societies.

The metaphysical part of the project required giving reasons to reject arguments against the autonomy of psychological-level causal claims which rely on the concepts of supervenience, physicalism, and causation. It was argued that a major premise in these arguments is the principle of the causal closure of the physical, but that once this principle is scrutinised closely, which it normally isn't, it does not provide the support for reductionism that many take it to.

It was argued that the correct account of the relationship between the cognitive and the physical levels is Virtual Machine Functionalism, as this not only shows how mental mechanisms can have the causal powers they do, through being built of physical parts, without being reducible to physical regularities, but also allows for a plausible story to be told of how those mechanisms evolved. This approach also makes it plausible to discover the kinds of organisation likely to lead to the emergence of minded creatures with subjective perspective on the world, this experiential aspect being a necessary property of being an embodied cognizer. With these parts in place, I believe we can justify a real distinction between literal and metaphorical uses of intentional descriptions of observed behaviour, which depends on whether the target of the description has been built, by evolution or design, with mechanisms that process information about the world which is used as the basis for actions to achieve the goals of the creature.

My goal has been to put cognitive kinds on a map that can be referred to on both scientific and philosophical journeys as a common reference point, where features of concern to both sets of interests are plotted. There are many parts that require further exploration, and some that will probably need to be redrawn, but I hope it serves as at least a roughly accurate depiction of the territory, one that helps to bring conceptual clarity to the field of cognitive science as it moves into a more mature phase that requires some settling on agreed definitions. The framework provided by the topographical account is one that will be filled in in practice by researchers in the various disciplines of cognitive science; its advantage is exactly that is allows pluralistic, interdisciplinary

practices to add to the accurate description of portions of the cognitive landscape without imposing a single overarching metaphysics. A problem in cognitive science, where multiple disciplines are searching for the right descriptions, is that all too often, a particular approach is touted as the one answer that can solve all the problems (e.g. computationalism, enactivism, eliminativism). Rather than cooperating, these research groups compete. The account promoted here provides the space for productive pluralism, not a relativistic pluralism that says all views are equally valid and equally trivial, but a pluralism where varied perspectives can be put together to form a complex vision.

This metaphysical framework is of a piece with Virtual Machine Functionalism, which likewise is a philosophical account with scientific implications. VMF has the resources to shape empirical research in investigating the 'how' of cognition, and to give direction to engineering projects in robotics and artificial intelligence research. Philosophically, it also helps rule on questions such as what counts as intelligent action. This is because it is not just output focused or behaviour centred, nor a matter of implementing an abstract algorithm. Unlike traditional functionalism, VMF is not a 'black box' theory; it does matter how the outcome is achieved, using real-time, real-world dynamics of the machine's physical parts. Understanding how the virtual machines that form the mind have evolved and developed, biologically and socially, provides the framework for interdisciplinary knowledge to be applied in explaining and predicting the behaviour of existing cognitive beings, as well as giving direction to engineers working on building cognitive machinery.

Finally, a further advantage of this view, with the class of intentional beings being seen as a distinguishable feature on the map of the multidimensional intensional similarity space defined by the sharing of identifiably similar features of virtual machine architecture, is that it allows a clear distinction to be made between metaphorical and literal uses of intentional explanations.

> And as he drove on, the rainclouds dragged down the sky after him, for, though he did not know it, Ron McKenna was a Rain God. All he knew was that his working days were miserable and he had a succession of lousy holidays. All the clouds knew was that they loved him and wanted to be near him, to cherish him, and to water him. (Adams, 1984, p. 15)

Bibliography

Adams, D., 1984. *So long, and thanks for all the fish.* London: Pan Books.

Adelson, E. H., 1995. *Checkershadow illusion.* [Online]
Available at: http://web.mit.edu/persci/people/adelson/checkershadow_illusion.html
[Accessed 15 January 2017].

Aleksander, I. & Dunmall, B., 2003. Axioms and Tests for the Presence of Minimal Consciousness in Agents. *Journal of Consciousness Studies,* 10(4–5), p. 7–18.

Armstrong, D., 1978. *A Theory of Universals.* Cambridge: Cambridge University Press.

Ayers, M. R., 1981. Locke Versus Aristotle on Natural Kinds. *The Journal of Philosophy,* 78(5), p. 247–272.

Baars, B. J., 1988. *A Cognitive Theory of Consciousness.* Cambridge, MA: Cambridge University Press.

Baars, B. J., 1997. *In the Theater of Consciousness.* Oxford: Oxford University Press.

Bach-y Rita, P. et al., 1969. Vision substitution by tactile image projection. *Nature ,* Issue 221, p. 963–964.

Bach-y-Rita, P. & Kercel, S. W., 2003. Sensory substitution and the human – machine interface. *Trends in Cognitive Sciences,* 7(12), pp. 541-546.

Baldwin, J. M., 1896. Consciousness and evolution. *Psychological Review,* Volume 3, p. 300–309.

Baron-Cohen, S., 1995. *Mindblindness: An Essay on Autism and Theory of Mind.* Cambridge, MA: MIT Press.

Barsalou, L., 2009. Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society of London: Biological Sciences,* Volume 364, p. 1281–1289.

Basso, D. &. O. B. M., 2006. The role of the feedforward paradigm in cognitive psychology. *Cognitive Processing,* 7(2), p. 73–88.

Berkeley, G., 1710. *Principles of Human Knowledge.* 2004 ed. London: Penguin.

Bermudez, J. L., 2003. Ascribing thoughts to non-linguistic creatures. *Facta Philosophica,* 5(2), pp. 313-34.

Bickhard, M., 2006. *The Dynamic nature of Representation.* [Online]
Available at: http://interdisciplines.org/adaptation/papers/2
[Accessed 3 February 2007].

Block, N., 1978. Troubles with Functionalism. *Minnesota Studies in the Philosophy of Science,* Volume 9, pp. 261-325.

Block, N., 1980. What is Functionalism?. In: N. Block, ed. *Readings in Philosophy of Psychology vol.1.* Cambridge, MA: Harvard University Press, pp. 171-184.

Boden, M. A., 2004. *The Creative Mind: Myths and Mechanisms.* London: Routledge.

Bourdieu, P., 1980. *The Logic of Practice.* Stanford: Stanford University Press.

Bourdieu, P., 2004. *Science of Science and Reflexivity.* Chicago: Polity Press and the University of Chicago.

Boyd, R., 1989. What Realism Implies and What it Does Not. *Dialectica,* 43(1-2), pp. 5-29.

Boyd, R., 1991. Realism, Anti-foundationalism, and the Enthusiasm for Natural kinds. *Philosophical Studies,* Volume 61, pp. 127-148.

Brandon, R. N., 2007. *Interdisciplines: Adaptation and Representation: The Theory of Biological Adaptation and Function.* [Online]
Available at: http://interdisciplines.org/adaptaion/papers/10/printable/paper
[Accessed 23 February 2007].

Brass, M. & Haggard, P., 2010. The hidden side of intentional action: the role of the anteriorinsular cortex. *Brain Structure and Function,* Volume 214, p. 603–610.

Broad, C., 1925. *The Mind and Its Place in Nature.* London: Routledge and Kegan Paul.

Brogaard, B., 2013. *The Brain Of The Real Rain Man.* [Online]
Available at: https://www.psychologytoday.com/blog/the-superhuman-mind/201303/the-brain-the-real-rain-man
[Accessed 19 January 2017].

Brooks, R., 1991. Intelligence without representation. *Artificial Intelligence,* 47(1-3), p. 139–159.

Burge, T., 1986. Individualism and Psychology. *Philosophical Review,* Issue 95, pp. 3-45.

Butterworth, G., 1995. An ecological perspective on the origins of self. In: J. Bermúdez, A. Marcel & N. Eilan, eds. *The Body and the Self.* Cambridge, MA: MIT/Bradford Press, p. 87–107.

Cantwell Smith, B., 1996. *On the Origin of Objects.* Cambridge MA: MIT Press.

Carnap, R., 1950. *Logical Foundations of Probability.* Chicago: University of Chicago Press.

Carroll, L., 1895. Chapter XI. In: *Sylvie and Bruno Concluded.* London: s.n.

Carruthers, P., 2002. The cognitive functions of language. *Behavioral and Brain Sciences,* Volume 25, pp. 657-726.

Cartwright, N., 1999. *The Dappled World: A Study of the Boundaries of Science.* Cambridge: Cambridge University Press.

Chalmers, D., 2002. Consciousness and Its Place in Nature. In: *Philosophy of Mind: Classical and Contemporary Readings.* s.l.:s.n.

Chalmers, D. J., 1995. Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies,* 2(3), pp. 200-219.

Chalmers, D. J., 1996. *The Conscious Mind: In Search of a Fundamental Theory.* New York and Oxford: Oxford University Press.

Chalmers, D. J., 2003. Consciousness and its Place in Nature. In: S. Stich & F. Warfield, eds. *Blackwell Guide to the Philosophy of Mind.* s.l.:Blackwell.

Chrisley, R., 1995. Non-conceptual content and robotics: Taking embodiment seriously. In: K. Ford, C. Glymour & P. Hayes, eds. *Android Epistemology.* Cambridge, MA: AAAI/MIT Press.

Chrisley, R., 2009. Synthetic Phenomenology. *International Journal of Machine Consciousness ,* 1(1), pp. 53-70.

Chrisley, R., 2010. Interactive Empiricism: The Philosopher in the Machine. In: K. Guy, ed. *Philosophy of Engineering: Volume 1 of the proceedings of a series of seminars held at The Royal Academy of Engineering.* London: The Royal Academy of Engineering, pp. 66-71.

Chrisley, R. & Sloman, A., 2016, in press. Functionalism, Revisionism and Qualia. *APA Newsletter on Philosophy and Computers,* 16(1).

Chrisley, R. & Sloman, A., 2016. Functionalism, Revisionism, and Qualia. *APA Newsletter: Philosophy and Computers*, Fall, 16(1), pp. 2-11.

Churchland, P. M., 1981. Eliminative Materialism and the Propositional Attitudes. *Journal of Philosophy,* Volume 78, p. 67–90.

Clark, A., 1994. Language of Thought (2). In: S. Guttenplan, ed. *A Companion to the Philosophy of Mind.* Oxford: Basil Blackwell.

Clark, A., 1997. *Being there: Putting brain, body and world together again.* Cambridge, MA: MIT.

Clark, A., 2002. Is Seeing All It Seems? Action, Reason and the Grand Illusion.. *Journal of Consciousness Studies,* 9(5-6), pp. 181-202.

Clark, A., 2008. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension.* s.l.:Oxford University Press.

Clark, A., 2016. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind.* Oxford: Oxford University Press.

Clark, A. & Chalmers, D., 1998. The Extended Mind. *Analysis,* 58(1), pp. 7-19.

Clark, A. & Grush, R., 1999. Towards a Cognitive Robotics. *Adaptive Behavior,* 7(1), pp. 5-16.

Clowes, R. W. & Chrisley, R., 2012. Virtualist Representation. *International Journal of Machine Consciousness,* 4(2), pp. 503-522.

Cooper, R., 2004. Why Hacking is wrong about human kinds. *British Journal for the Philosophy of Science,* Issue 55, pp. 73-85.

Cooper, R., 2005. *Classifying Madness: A Philosophical Examination of the Diagnostic and Statistical Manual of Mental Disorders.* s.l.:Springer Netherlands.

Cotterill, R., 2003. CyberChild: A Simulation Test-Bed for Consciousness Studies. *Journal of Consciousness Studies,* 10(4-5), pp. 31-45.

Cussins, A., 1990. The Connectionist Construction of Concepts. In: M. Boden, ed. *The Philosophy of Artificial Intelligence.* s.l.:Oxford University Press, pp. 368-440.

Daly, C. J., 2008. Fictionalism and the attitudes. *Philosophical Studies,* 139(3), p. 423–440.

Darley, J. M. & Latané, B., 1968. Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology,* Issue 8, p. 377–383.

Davidson, D., 1966. Emeroses by other names. *Journal of Philosophy,* Volume 63, pp. 778-80.

Davidson, D., 1969. The Individuation of Events. In: N. Rescher, ed. *Essays in Honor of Carl G. Hempel.* Dordrecht: D. Reidel, pp. 295-309.

Davidson, D., 1970. Mental Events. In: L. Foster & J. Swanson, eds. *Experience and Theory .* London: Duckworth.

Davidson, D., 1980. *Essays on Actions and Events.* Oxford: Clarendon Press.

Davidson, D., 1984. Thought and Talk. In: *Inquiries into Truth and Interpretation.* Oxford: Clarendon Press, pp. 155-179.

Davidson, D., 1985. Rational Animals. In: E. Lepore & B. McLaughlin, eds. *Actions and Events: Perspectives on the Philosophy of Donald Davidson.* New York: Basil Blackwell.

Dayan, P., 2008. The Role of Value Systems in Decision Making. In: E. C & S. W, eds. *Strüngmann Forum Report: Better Than Conscious? Decision Making, the Human Mind, and Implications For Institutions.* Frankfurt, Germany: MIT Press, pp. 51-70.

Dennett, D. C., 1981. *Brainstorms: Philosophical Essays on Mind and Psychology.* Cambridge, MA: MIT Press.

Dennett, D. C., 1987. *The Intentional Stance.* Cambridge, MA: MIT Press.

Dennett, D. C., 1991. *Consciousness Explained.* London: The Penguin Press.

Dennett, D. C., 1992. The Self as a Center of Narrative Gravity. In: F. Kessel, P. Cole & D. Johnson, eds. *Self and Consciousness: Multiple Perspectives.* Hillsdale, NJ: Erlbaum.

Dennett, D. C., 1996. *Kinds of Minds: Toward an Understanding of Consciousness.* New York: Basic Books, Inc..

Dennett, D. C., 2003. *Freedom Evolves.* New York: Viking Press.

Dennett, D. C., 2005. *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness.* s.l.:MIT.

Devitt, M., 1981. *Designation.* New York: Columbia University Press.

Dupré, J., 1993. *The Disorder of Things : Metaphysical Foundations of the Disunity of Science.* Cambridge MA: Harvard University Press.

Dupré, J., 1996. Promiscuous Realism: Reply to Wilson. *The British Journal for the Philosophy of Science,* 47(3), pp. 441-444.

Edelman, G. M., 1989. *The Remembered Present.* New York: Basic Books.

Evans, G., 1982. *The Varieties of Reference.* Oxford: Oxford University Press.

Fischer, J. M. & Ravizza, M., 1998. *Responsibility and Control: A Theory of Moral Responsibility.* Oxford: Oxford University Press.

Fodor, J. A., 1974. Special Sciences (Or: the Disunity of Science as a Working Hypothesis). *Synthese,* 28(2), pp. 97-115.

Fodor, J. A., 1975. *The Language of Thought.* Cambridge, MA: Harvard University Press.

Fodor, J. A., 1980. Methodological Solipsism Considered as a Research Strategy in Cognitive Science. *Behavioral and Brain Sciences,* 3(1), pp. 63-73.

Fodor, J. A., 1983. *The Modularity of Mind.* Cambridge, MA: MIT Press.

Fodor, J. A., 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind.* Cambridge, MA: MIT Press.

Fodor, J. A., 1994. *The Elm and the Expert: Mentalese and Its Semantics.* Cambridge, MA: MIT Press.

Fodor, J. A., 1997. Special Sciences: Still Autonomous After All These Years. *Philosophical Perspectives,* 31(11), pp. 149-163.

Fodor, J. A., 2008. *LOT 2: The Language of Thought Revisited.* Oxford: Oxford University Press.

Fodor, J. & Pylyshyn, Z., 1988. Connectionism and Cognitive Architecture: a Critical Analysis. *Cognition,* Volume 28, pp. 3-71.

Franklin, S., 2003. IDA: A Conscious Artifact?. *Journal of Consciousness Studies,* 10(4-5), p. 47–66.

Gallagher, S., 2001. The Practice of Mind: Theory, Simulation or Primary Interaction?. *Journal of Consciousness Studies,* 8(5–7), p. 83–108.

Gallese, V., 2001. The 'Shared Manifold' Hypothesis: From Mirror Neurons To Empathy. *Journal of Consciousness Studies,* 8(5–7), p. 33–50.

Gärdenfors, P., 2000. *Conceptual spaces: The geometry of thought.* Cambridge, MA: MIT Press.

Gazzinga, M., 1970. *The Bisected Brain.* New York: Appleton-Century-Croft.

Gibson, J., 1979. *The Ecological Approach to Visual Perception.* Boston: Houghton Mifflin.

Godfrey-Smith, P., 2003. Folk Psychology Under Stress: Comments on Susan Hurley's 'Animal Action in the Space of Reasons'. *Mind and Language,* 10(3), pp. 266-272.

Goff, P., 2009. Why Panpsychism doesn't Help Us Explain Consciousness. *dialectica,* 63(3), pp. 289-311.

Goodale, M. A. & Milner, A. D., 1992. Separate visual pathways for perception and action. *Trends in Neurosciences,* Issue 15, pp. 97-112.

Goodale, M. A., Milner, A. D., Jakobson, L. S. & Carey, D. P., 1991. A neurological dissociation between perceiving objects and grasping them. *Nature,* Issue 349, pp. 154-156.

Goodman, N., 1955. *Fact, Fiction and Forecast.* Cambridge: Harvard University Press.

Gopnik, A. & Meltzoff, A., 1997. *Words, Thoughts, and Theories.* Cambridge, MA: MIT Press.

Gould, S. J. & Lewontin, R. C., 1979. The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adptationist Programme. *Procedings of the Royal Society of London, Series B,* 205(1161), pp. 581-598.

Hacking, I., 1990. Natural Kinds. In: R. Barret & R. Gibson, eds. *Perspectives on Quine.* Oxford: Blackwell, pp. 129-43.

Hacking, I., 1991. A Tradition of Natural Kinds. *Philosophical Studies,* Volume 61, pp. 109-126.

Hameroff, S. R. & Penrose, R., 1996. Conscious events as orchestrated spacetime selections. *Journal of Consciousness Studies,* Issue 3, pp. 36-53.

Hare, R., 1952. *The Language of Morals.* Oxford: Clarendon Press.

Haugeland, J., 1982. Weak Supervenience. *American Philosophical Quarterly,* Issue 19, p. 93–101.

Heal, J., 2005. Joint attention and understanding the mind. In: J. Roessler, ed. *Joint attention: Communication and other minds.* Oxford: Oxford University Press, pp. 34-44.

Heidegger, M., 1927. *Being and time.* New York: Harper & Row.

Heil, J. & Robb, D., 2003. Mental properties. *American Philosophical Quarterly,* 40(3), pp. 175-196.

Heyes, C. & Dickinson, A., 1990. The Intentionality of Animal Action. *Mind and Language,* Volume 5, pp. 87-104.

Holland, O. & Goodman, R., 2003. Robots With Internal Models: A Route to Machine Consciousness?. *Journal of Consciousness Studies,* 10(4–5), p. pp. 77–109.

Honderich, T., 2006. Radical externalism. *Journal of Consciousness Studies,* 13(7-8), pp. 3-13.

Horgan, T., 1982. Supervenience and Microphysics. *Pacific Philosophical Quarterly,* Issue 63, pp. 29-43.

Horgan, T., 1993. From Supervenience to Superdupervenience: Meeting the Demands of a Material World. *Mind,* 102(408), p. 555–586.

Hull, D., 1972. Reductionism in genetics – biology or philosophy?. *Philosophy of Science,* Volume 39, p. 491–499.

Hume, D., 1738. *A Treatise of Human Nature.* s.l.:s.n.

Humphreys, P., 1997a. Emergence, Not Supervenience. *Philosophy of Science,* 64(Supplement, Proceedings of the 1996 Biennial Meetings of the Philosophy of Science Association. Part II: Symposia Papers), pp. S337-S345.

Humphreys, P., 1997b. How Properties Emerge. *Philosophy of Science,* 64(1), pp. 1-17.

Hurley, S., 2003. Animal Action in the Space of Reasons. *Mind and Language,* 18(3), pp. 231-256.

Hurley, S., 2006. Making Sense of Animals. In: H. Susan & M. Nudds, eds. *Rational Animals.* Oxford: Oxfor University Press, pp. 139-174.

Hurley, S., 2006. Varieties of externalism. In: R. Menary, ed. *The Extended Mind.* s.l.:Ashgate, pp. 1-35.

Hutto, D., 2013. Fictionalism about folk psychology. *The Monist,* 96(4), pp. 582-604.

Israel, R., 2004. Two interpretations of 'grue' – or how to misunderstand the new riddle of induction. *Analysis 64.4, October 2004, pp. 335–39.,* 64(4), pp. 335-39.

Israel, R., 2006. Projectibility and Explainability or How to Draw a New Picture of Inductive Practices. *Journal for General Philosophy of Science,* Volume 37, pp. 269-286.

Ivanov, I., 2016. *Observational concepts and experience.* PhD Thesis: University of Warwick.

Jackson, F., 1982. Epiphenomenal Qualia. *Philosophical Quarterly,* Issue 32, p. 127–136.

Jackson, F., 1998. *From Metaphysics to Ethics: A Defense of Conceptual Analysis.* Oxford: Clarendon.

James, W., 1890. *The Principles of Psychology.* Cambridge, MA: Harvard University Press.

Kim, J., 1976. Events as Property Exemplifications. In: M. Brand & D. Walton, eds. *Action Theory.* s.l.:s.n., pp. 310-326.

Kim, J., 1990. Supervenience as a Philosophical Concept. *Metaphilosophy,* Volume 21, pp. 1-27.

Kim, J., 1992. 'Downward Causation' in Emergentism and Nonreductive Physicalism. In: F. K. Beckermann, ed. *Emergence or Reduction?.* Berlin: de Gruyter, pp. 119-138.

Kim, J., 1992. Multiple Realization and the Metaphysics of Reduction. *Philosophy and Phenomenological Research,* 52(1), pp. 1-26.

Kim, J., 1996. *Philosophy of Mind.* Boulder, Colorado: Westview Press.

Kim, J., 2005. *Physicalism, or Something Near Enough.* s.l.:Princeton University Press.

Kripke, S., 1972. *Naming & Necessity.* s.l.:Blackwell.

Kripke, S., 1982. *Wittgenstein on Rules and Private Language.* Cambridge, MA: Harvard University Press.

Kuhn, T. S., 1962. *The Structure of Scientific Revolutions.* Chicago: University of Chicago Press.

Kuhn, T. S., 2000. *The Road since Structure.* Chicago: Chicago University Press.

Lakoff, G. & Johnson, M., 1999. *Philosophy In The Flesh: The Embodied Mind and Its Challenge to Western Thought.* s.l.:Basic Books.

LaPorte, J., 2000. Rigidity and Kind. *Philospohical Studies,* Volume 97, pp. 263-316.

LaPorte, J., 2004. *Natural Kinds and Conceptual Change.* s.l.:Cambridge University Press.

Levine, J., 1983. Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly,* Issue 64, pp. 354-361.

Lewis, C. I., 1929. *Mind and the World Order: Outline of a Theory of Knowledge.* New York: Charles Scribners.

Lewis, D., 1966. An Argument for the Identity Theory. *Journal of Philosophy,* Volume 63, p. 17–25.

Lewis, D., 1973. Causation. *Journal of Philosophy,* Volume 70, p. 556–567.

Lewontin, R. C., 1970. The Units of Selection. *Annual Review of Ecology and Systematics,* Volume 1, pp. 1-18.

Libet, B., 1985. Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences,* Issue 8, pp. 529-566.

Libet, B., 1999. Do We Have Free Will?. *Journal of Consciousness Studies,* 6(8-9), pp. 47-57.

List, C. & Menzies, P., 2014. The exclusion argument against free will, and what's wrong with it. In: H. Beebee, C. Hitchcock & H. Price, eds. *Making a Difference.* Oxford: Oxford University Press.

Locke, J., 1700. *An Essay Concerning Human Understanding.* s.l.:Oxford University Press.

Lowe, E. J., 2000. Causal Closure Principles and Emergentism. *Philosophy,* Volume 75, pp. 571-585.

Lowe, E. J., 2008a. Free Agency, Causation and Action Explanation. In: C. Sandis, ed. *New Essays on the Explanation of Action.* s.l.:Palgrave Macmillan.

Lowe, E. J., 2008b. *Personal Agency: The Metaphysics of Mind and Action.* Oxford: Oxford University Press.

Machery, E., 2003. Concepts Are Not a Natural Kind. *Philosophy of Science,* Issue 72, pp. 444-467.

Mackie, J. L., 1974. *The Cement of the Universe.* Oxford: Oxford University Press.

Malcolm, N., 1972 -1973. Thoughtless Brutes. *Proceedings and Addresses of the American Philosophical Association,* Volume 46, pp. 5-20.

Mameli, M., 2001. Mindreading, Mindshaping, and Evolution. *Biology and Philosophy,* Volume 16, pp. 597-628.

Mameli, M., 2004. Nongenetic selection and nongenetic inheritance. *British Journal for the Philosophy of Science,* 55(1), pp. 35-71.

Mandik, P. & Clark, A., 2002. Selective Representing and World-Making. *Minds and Machines,* Issue 12, pp. 383-395.

Manzotti, R., 2006. A Process Oriented View of Conscious Perception. *Journal of Consciousness Studies,* 13(6), pp. 7-41.

Margolis, E. & Laurence, S., 1999. *Concepts: Core Readings.* Cambridge, MA: MIT Press.

Martin, M. G. F., 2007. *School of advanced Study, University of London.* [Online]
Available at: http://sas-space.sas.ac.uk/630/1/M_Martin_Alienated.pdf
[Accessed 22 March 2013].

McDowell, J., 1994. The Content of Perceptual Experience. *The Philosophical Quarterly,* 44(175), pp. 190-205.

McGinn, C., 1978. Mental States, Natural Kinds and Psychophysical Laws. *Proceedings of the Aristotelian Society, Supplementary Volumes,* Volume 52, pp. 195-236.

McGinn, C., 1993. *Problems in Philosophy: The Limits of Inquiry.* Oxford: Blackwell.

McGinn, C., 2006. Hard Questions: comments on Galen Strawson. *Journal of Consciousness Studies,* 13(10-11), pp. 90-99.

McGurk, H. & MacDonald, J., 1976. Hearing lips and seeing voices. *Nature*, 23-30 December, pp. 746-748.

McLaughlin, B., 1995. Varieties of supervenience. In: E. Savellos & U. Yalcin, eds. *Supervenience: New Essays.* Cambridge: Cambridge University Press, pp. 16-59.

McLaughlin, B. a. B. K., 2008. *Supervenience.* [Online]
Available at: http://plato.stanford.edu/archives/fall2008/entries/supervenience/
[Accessed 20 July 2009].

McLaughlin, B. P., 1984. Perception, Causation, and Supervenience. *Midwest Studies in Philosophy,* Volume 9, p. 569–92.

Mead, G. H., 1934. *Mind, Self, and Society: From the Standpoint of a Social Behaviorist.* Chicago: University of Chicago Press.

Mellor, D. H., 1977. Natural Kinds. *The British Journal for the Philosophy of Science,* 28(4), pp. 299-312.

Merleau-Ponty, M., 1945. *Phenomenology of Perception.* Paris: Gallimard.

Metzinger, T., 2003. *Being No One: The Self-Model Theory of Subjectivity.* Cambridge, MA: MIT Press.

Millikan, R. G., 1999. Historical Kinds and the “Special Sciences”. *Philosophical Studies,* Volume 95, pp. 45-65.

Mill, J. S., 1843. *A System of Logic.* London: John W Parker.

Milner, A. D. & Goodale, M. A., 1995. *The Visual Brain in Action.* Oxford: Oxford University Press.

Minsky, M., 1986. *The Society of Mind.* New York: Simon & Schuster.

Morgan, C. L., 1894. *An introduction to comparative psychology.* London: W. Scott.

Morgan, L., 1923. *Emergent Evolution.* London: Williams & Norgate.

Nagel, E., 1961. *The Structure of Science: Problems in the Logic of Scientific Explanation.* New York: Harcourt, Brace, and World.

Nagel, T., 1971. Brain Bisection and the Unity of Consciousness. *Synthese,* 22(3/4), pp. 396-413.

Nagel, T., 1974. What is it like to be a bat?. *Philosophical Review,* Issue 83, p. 435–50.

Nagel, T., 1979. Panpsychism. In: T. Nagel, ed. *Mortal Questions.* Cambridge: Cambridge University Press, pp. 181-195.

Ney, A., 2016. Microphysical Causation and the Case for Physicalism. *Analytic Philosophy,* 57(2), pp. 141-164.

Noë, A., 2004. *Action in perception.* Cambridge, Mass : MIT Press.

Noë, A., 2006. Experience Without the Head. In: T. S. Gendler & J. Hawthorne, eds. *Perceptual Experience.* s.l.:Oxford University Press.

Nunberg, G., 2006. *Last Planet Standing.* [Online]
Available at: http://people.ischool.berkeley.edu/~nunberg/pluto.html
[Accessed 17 July 2016].

O'Regan, J., 1992. Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology/Revue Canadienne de Psychologie,* 46(3), pp. 461-488.

O'Regan, J. K. & Noë, A., 2001. A sensorimotor account of vision and visual consciousness. *Behavioural and Brain Sciences,* 24(5), pp. 939-1031.

Panksepp, J., 1998. *Affective neuroscience: The foundations of human and animal emotions.* New York: Oxford University Press.

Papineau, D., 2006. Comments on Galen Strawson. *Journal of Consciousness Studies,* 13(10-11), pp. 100-109.

Papineau, D., 2010. Can any sciences be special?. In: C. Macdonald & G. Macdonald, eds. *Emergence in Mind.* Oxford: Oxford University Press, pp. 179-19.

Pascual-Leone, A. & Walsh, V., 2001. Fast Backprojections from the Motion to the Primary Visual Area Necessary for Visual Awareness. *Science,* Issue 292, pp. 510-512.

Pietroski, P., 2000. *Causing Actions.* New York: Oxford University Press.

Plato, 1997. *Complete works.* Indianapolis: Hackett Publishing Company.

Poiriera, C., De Voldera, A. G. & Scheiber, C., 2007. What neuroimaging tells us about sensory substitution. *Neuroscience and Biobehavioral Reviews ,* Issue 31, p. 1064–1070.

Prinz, J. J., 2004. *Furnishing the Mind: Concepts and Their Perceptual Basis.* s.l.:MIT Press.

Prinz, J. J., 2006. Putting the Brakes on Enactive Perception. *PSYCHE,* 12(1), pp. 1-19.

Putnam, H., 1969. On Properties. In: J. Kim & E. Sosa, eds. *Metaphysics .* s.l.:Blackwell, pp. 243-52.

Putnam, H., 1975. The Meaning of Meaning. *Minnesota Studies in the Philosophy of Science,* Volume 7, pp. 131-193.

Putnam, H., 1981. *Reason, Truth and History.* Cambridge: Cambridge University Press.

Quine, W. V., 1969. Natural Kinds. In: J. Kim & E. Sosa, eds. *Metaphysics.* s.l.:Blackwell, pp. 233-42.

Richter, H. et al., 2002. Long-term adaptation to prism-induced inversion of the retinal images. *Experimental Brain Research,* 144(4), p. 445–457.

Rockwell, W. T., 2007. *Neither Brain nor Ghost: A Nondualist Alternative to the Mind-Brain Identity Theory.* Cambridge, MA: MIT Press.

Rosch, E., 1999. Principles of Categorization. In: E. Margolis & S. Laurence, eds. *Concepts: Core Readings.* Cambridge, MA: MIT Press, pp. 189-206.

Rowlands, M., 2002. Two dogmas of consciousness. *Journal of Consciousness Studies,* 9(5-6), pp. 158-180.

Ruse, M., 1987. Biological Species: Natural Kinds, Individuals, or What?. *The British ournal for the Philosophy of Science,* 38(2), pp. 225-242.

Russell, B., 1905. On Denoting. *Mind,* Volume 14, p. 479–493.

Russell, B., 1927. *The Analysis of Matter.* London: Routledge.

Russell, B., 1948. *Human Knowledge: Its Scope and Limits.* London: George Allen & Unwin.

Ryle, G., 1946. Knowing How and Knowing That. In: *Collected Papers (Volume 2).* New York: Barnes and Nobles, p. 212–25.

Ryle, G., 1949. *The Concept of Mind.* London: Hutchinson.

Salmon, W., 1984. *Scientific Explanation and the Causal Structure of the World.* Princeton: Princeton University Press.

Schwartz, S. P., 2002. Kinds, General Terms, and Rigidity: A Reply to LaPorte. *Philosophical Studies,* Volume 109, pp. 265-277.

Searle, J. R., 1980. Minds, brains, and programs. *Behavioral and Brain Sciences,* 3(3), pp. 417-457.

Searle, J. R., 1992. *The Rediscovery of the Mind.* Cambridge, MA: MIT Press.

Silberstein, M., 2002. Reduction, Emergence and Explanation. In: P. Machamer & M. Silberstein, eds. *The Blackwell Guide to the Philosophy of Science.* Oxford: Blackwell, pp. 80-108.

Simons, D. J. & Levin, D. T., 1997. Change blindness. *Trends in Cognitive Sciences,* 1(7), p. 261–267.

Skow, B., 2014. Are There Non-causal Explanations (of Particular Events)?. *British Journal for the Philosophy of Science,* Volume 65, pp. 445-67.

Sloman, A., 1987. Motives, mechanisms, and emotions. *Cognition and Emotion,* 1(3), pp. 217-233.

Sloman, A., 1992. The Emperor's Real Mind (Review of: Roger Penrose: The Emperor's New Mind: Concerning Computers, Minds and the Laws of Physics. *Artificial Intelligence,* Issue 56, pp. 355-396.

Sloman, A., 2007. *Why Some Machines May Need Qualia and How They Can Have Them: Including a Demanding New Turing Test for Robot Philosophers.* [Online]
Available at: http://www.cs.bham.ac.uk/~axs
[Accessed 21 June 2011].

Sloman, A., 2008. The well-designed young mathematician. *Artificial Intelligence ,* 172(18), pp. 2015-2034.

Sloman, A., 2011a. *Virtual Machinery and Evolution of Mind (Part 1).* [Online]
Available at: http://www.cs.bham.ac.uk/~axs)
[Accessed 21 June 2011].

Sloman, A., 2011b. *Virtual Machinery and Evolution of Mind (Part 2).* [Online]
Available at: http://www.cs.bham.ac.uk/~axs)
[Accessed 21 June 2011].

Sloman, A., 2013. *Virtual Machine Functionalism (VMF) (The only form of functionalism worth taking seriously in Philosophy of Mind and theories of Consciousness).* [Online]
Available at: http://www.cs.bham.ac.uk/research/projects/cogaff/misc/vm-functionalism.html#block95
[Accessed 14 January 2017].

Sloman, A. & Chrisley, R., 2003. Virtual Machines and Consciousness. *Journal of Consciousness Studies,* 10(4-5), pp. 133-72.

Sloman, A., Chrisley, R. & Sheutz, M., 2003. The Architectural Basis of Affective States and Processes. In: Fellous & Arbib, eds. *Who Needs Emotions?: The Brain Meets the Machine.* s.l.:Oxford University Press.

Smith, E. E., 1999. The Exemplar View. In: E. Margolis & S. Laurence, eds. *Concepts: Core Readings.* Cambridge, MA: MIT Press, pp. 207-221.

Soon, C. S., Brass, M., Heinze, H.-J. & Haynes, J. D., 2008. Unconscious determinants of free decisions in the human brain. *Nature Neuroscience,* 11(5), pp. 543-545.

Steels, L., 2003. Language Re-Entrance and the 'Inner Voice'. *Journal of Consciousness Studies,* 10(4-5), p. 173–85.

Sterelney, K., 2003. Charting Control-Space: Comments on Susan Hurley's 'Animal Action in the Space of Reasons'. *Mind and Language,* 18(3), pp. 257-166.

Stoljar, D., 2009. *Physicalism,* s.l.: s.n.

Strawson, G., 2006. Realistic Monism: Why Physicalism Entails Panpsychism. *Journal of Consciousness Studies,* 13(10-11), pp. 3-31.

Stuss, D. T., 1991. Self, Awareness, and the Frontal Lobes: A Neuropsychological Perspective. In: S. J & G. G. R, eds. *The Self: Interdisciplinary Approaches.* New York: Springer-Verlag, p. 255–278.

Swinburne, R. G., 1968. Grue. *Analysis,* 28(4), pp. 123-128.

The Independent, 2006. Solar system welcomes three new planets. *The Independent*, 15 August.

Thompson, E., 2008. Representationalism and the phenomenology of mental imagery. *Synthese,* 160(3), pp. 203-213.

Tomasello, M., 1999. *The Cultural Origins of Human Cognition.* Cambridge, MA: Harvard University Press.

Ünver, E. & Güntürkün, O., 2014. Evidence for interhemispheric conflictduring meta-control in pigeons. *Behavioural Brain Research,* Issue 270, pp. 146-150.

Van Gulick, R., 2001. Reduction, Emergence and Other Recent Options on the Mind/Body Problem. *Journal of Consciousness Studies,* 8(9-10), pp. 1-34.

Varela, F. J., Thompson, E. & Rosch, E., 1991. *The embodied mind: Cognitive science and human experience.* Cambridge, MA: MIT Press.

Voss, S., 2004. *The Place of Mind in Nature.* [Online]
Available at: http://www.phil.boun.edu.tr/files/voss/voss_paper2.html
[Accessed 2005].

Vygotsky, L., 1978. *Mind in Society: The Development of Higher Psychological Processes.* Cambridge, MA: Harvard University Press.

Wason, P. C., 1968. Reasoning about a rule. *Quarterly Journal of Experimental Psychology,* 20(3), pp. 273-281.

Wilson, M., 2002. Six views of embodied cognition. *Psychonomic Bulletin and Review,* Volume 9, pp. 625-636.

Wittgenstein, L., 1953. *Philosophical Investigations.* Oxford: Basil Blackwell Ltd..

Wolman, D., 2012. The split brain: A tale of two halves. *Nature*, 14 March, pp. 260-263.

Zahavi, D., 1994. Husserl's Phenomenology of the Body. *Études Phénoménologiques,* Issue 19, p. 63–84.

Zaidel, E., 1994. Interhemispheric transfer in the split brain: Long term status following complete cerebral commissurotomy. In: R. H. Davidson & K. Hugdahl, eds. *Human Laterality.* Cambridge, MA: MIT Press , pp. 491-532.