Categorizing the Mental

Eric Hochstein

Forthcoming in *The Philosophical Quarterly*

Abstract: A common view in the philosophy of mind and philosophy of psychology is that there is an ideally correct way of categorizing the structures and operations of the mind, and that the goal of neuroscience and psychology is to find this correct categorizational scheme. Categories which cannot find a place within this correct framework ought to be eliminated from scientific practice. In this paper, I argue that this general idea runs counter to productive scientific practices. Such a view ignores the plurality of aims and goals that neuroscientists and psychologists have in studying mental phenomena, and the necessity of employing distinct classificatory frameworks to achieve them.

In his 2011 paper 'Resisting "Weakness of Will"', Neil Levy argues that we ought to question the scientific value of the concept of 'weakness of will' in domains like psychology and neuroscience on the grounds that…

> ...weakness of the will, as a folk psychological concept, does not correspond to a psychological kind. In postulating its existence, we do not cut our cognitive nature at its joints. The effects of ego depletion are not limited to self-control nor are its causes: if

> we wish to be able to explain our failures of practical rationality, we should abandon the notion. (Levi 2011: 152)

The motivation behind Levi's argument is that if ‘weakness of will’ does not correspond to a distinct mental process or phenomenon, but is instead an ill-defined category for a number of disparate cognitive processes, then the category fails to do any explanatory work, and ought to be replaced. Implicit in Levi's argument is the idea that there is an ideally correct way of categorizing the cognitive operations and processes of the mind (one which ‘cuts our cognitive nature at its joints’), and that the goal of neuroscience and psychology is to develop such a categorizational scheme. Categories which cannot find a place within this correct framework ought to be eliminated from scientific practice.

This type of argument can be found throughout the philosophy of mind and philosophy of psychology. Some argue that this ideally correct categorizational scheme will be one that accurately describes the cognitive structures and operations of the mind, while others argue that it will be one that accurately denotes natural kinds (for some examples, see: Churchland 1981; Stich 1983, 1996; Devitt & Sterelny 1987: 242; Churchland 1989; Sehon 1997; Bickle 1998, 2003, 2006; Griffiths 2004, 2007; Murphy 2006; Machery 2009; Levy 2011; Penn 2012). In this paper, I argue that the assumption that there is a single ideally correct way of classifying or categorizing mental phenomena, and that neuroscience and psychology should adhere to this correct scheme, runs counter to productive scientific practices in these domains. Such a view largely ignores or glosses over the plurality of aims and goals that neuroscientists and psychologists have in studying mental phenomena. The best classificatory scheme to use for some of these goals will not be the same as the one best suited for others. Different aims and goals will be facilitated by different categorizational schemes, and thus there will not be a single classificatory scheme that will be useful for all of them.

In order to demonstrate this, I begin in Section 1 by focusing on two specific examples of this style of argument to better illustrate the intuitions underlying it. Next, in section two, I defend the claim that different scientific aims and goals are facilitated by the use of distinct ways of classifying and describing systems. In section three, I turn to the study of the mind in particular and demonstrate that models which categorize and classify the mind in a myriad of different ways are both commonplace and important to scientific practice in neuroscience and psychology.

1. Science and the Pursuit of the Correct Categorizational Scheme

The idea that there is a single correct way of classifying mental phenomena, and that science is primarily in pursuit of this classification, is an extremely common view in the philosophical literature. Take, for instance, arguments against the continued usage of propositional attitude ascriptions in science:

> Consider, for instance, the argument for the elimination of beliefs, desires, and other propositional attitudes (Churchland 1981; Stich 1983, 1996). It is argued that "belief" is defined by the role of the concept of belief in the generalizations about beliefs that we hold explicitly or implicitly. It is then argued that scientific evidence, for example,

> evidence drawn from neuropsychology (Churchland 1981) or from artificial intelligence (Ramsey *et al.* 1990), shows that no entities fulfill the role that defines “belief”. It is inferred that “belief” fails to refer, hence, that beliefs do not exist. (Machery 2009: 224)

What makes propositional attitude ascriptions inappropriate as part of scientific methodology according to these positions is that their terms *fail to refer*. In other words, there are no structures in the brain that have the sort of causal properties or roles that intentional psychology presupposes. Thus, these terms do not properly characterize the neurological/cognitive mechanisms that generate behaviour, and thus ought to be removed from our scientific vocabulary. The underlying assumption is that if there are distinct sets of cognitive structures and processes that generate mental phenomena, then it is the goal of science to correctly categorize and identify these structures and processes. If propositional attitudes cannot find a place within this correct classificatory scheme, then they ought to be removed from scientific practice.

While this style of argument is implicit in much of contemporary philosophy of mind, it would be helpful for our purposes to focus our attention on two particular examples in more detail to help flesh out the intuitions that underlie them. With this in mind, let us turn our attention to Paul Griffiths’ argument for the scientific elimination of the category ‘emotion’ (1997, 2004), as well as Edouard Machery's argument for the scientific elimination of the category ‘concept’ (2009).

1.1. Griffiths and the Category of ‘Emotion’

According to Paul Griffiths, the category of ‘emotion’ as used in psychology does not neatly map onto, or cleave along the lines of, any uniform set of psychological mechanisms or processes. Instead, the term partially refers to numerous different and unrelated cognitive mechanisms. In virtue of this, he proposes that the term is scientifically unhelpful, and that a correct categorization of the mind is one that does not subsume different types of mechanisms and processes under a single ill-fitting category. In his words:

> In earlier work I have described my position as a form of eliminitivism about emotion, because it implies that the term “emotion” and some specific emotion terms like “anger” are examples of “partial reference”. The term “jade” is the standard example of partial reference. “Jade” may be either of two different stones, jadeite or nephrite, and the term “jade” partially refers to each of these two kinds of stone. It follows from this fact that for the purposes of geology or chemistry, jade cannot be treated as a single kind of thing. The properties of the two minerals have to be investigated separately, their geological origins explained separately, and their abundance in unexplored geological deposits predicted separately. In a similar fashion, the sciences of the mind will have to develop separate theories of the various different kinds of emotion and also of the various different kinds of some particular emotions. In the same sense that there is really no such thing as jade, only jadeite and nephrite, there is no such thing as emotion, only affect programs, domain-specific biases in motivation, socially sustained pretences, and other more specific categories of psychological state and process that

> have been identified or hypothesized in the varied literature that sets out to address human emotion. (Griffiths 2004: 901-2)

Griffiths believes that just as the category 'jade' is scientifically inappropriate in geology in virtue of ignoring structural differences between jadeite and nephrite, so too is the category 'emotion' scientifically inappropriate for ignoring structural and causal differences between the various individual emotions.

Notice that a key part of Griffiths' argument is the idea that jade 'cannot be treated as a single kind of thing' for the purposes of geology or chemistry because the correct classificatory scheme for these domains to use is one that does not involve categories that partially refer to different types of stones. The implication being that for any given domain, there is a single correct way of categorizing systems; one which respects the structural and casual differences between systems (and thus avoids cases of partial reference). A proper scientific account is one which adheres to this correct categorizational scheme. Thus, if we have reasons for believing that a certain category (e.g. 'jade' in geology, 'emotion' in psychology) does not have a place in this correct conceptual framework, then it is unlikely to prove useful to that scientific domain.

It is worth noting that Griffiths is happy to grant that different domains of science each have their own ideal categorizational scheme (and thus he does not appear to be committed to the truth or falsity of theory reduction between scientific domains). However, he insists that for *any particular scientific domain* (e.g. psychology), the interests and goals of that domain dictate that there is a single correct way of parsing systems and phenomena. He says:

> The category of domestic pets is not a good category for investigating morphology, physiology, or behavior, but might be a natural category in some social psychological theory or, of course, in a theory about domestication. Similarly, emotion is not a natural kind relative to the domains of properties that are the focus of investigation in psychology and the neurosciences, or so I have argued. (2004: 905)

Notice that according to this view, the correct categorizational scheme for the sciences of the mind (or, as he puts it, ‘psychology and the neurosciences’) is not one that talks about ‘emotions’ because the focus of investigation for this domain dictates the correct categorizational scheme for it to employ, and ‘emotion’ is an instance of partial reference within this framework. In this respect, the value of the category is dependent on whether it fits with whatever the best categorizational scheme is thought to be for the domain of psychology and the neurosciences.

1.2. Machery and the Category of ‘Concept’

Machery's argument for the elimination of the category of ‘concept’ from psychology and neuroscience runs parallel to Griffiths’ argument regarding ‘emotion’. According to Machery, the category ‘concept’, as traditionally used by psychologists, does not refer to anything that constitutes a natural kind, and thus ought to be eliminated from scientific methodology as a result. According to Machery:

> A natural kind is a class about which many generalizations can be formulated: its members tend to have many properties in common. These generalizations are not accidental: there is at least one causal mechanism that explains why its members tend to have those properties. (2009: 232-3)

For Machery, what defines a natural kind is that all the tokens of a given kind have the same relevant properties in virtue of being the product of the same causal mechanism. Thus, his conception of natural kinds is rooted in a story about structured physical states and their causal processes. And for Machery, there is a single correct way of classifying such structures and processes. Deviation from this categorizational scheme is, according to Machery, grounds for scientific elimination:

> If "concept" does not pick out a natural kind, then it is unlikely to be a useful notion in psychology. It is even likely to stand in the way of progress in psychology, by preventing the development of a more adequate classificatory scheme that would identify the relevant natural kinds. If this is the case, the term "concept" ought to be eliminated from the theoretical vocabulary of psychology and replaced with more adequate theoretical terms. (2009: 230)

Here we see quite clearly the suggestion that the only good scientific methodologies are those that conform to our correct categorization of the system's underlying structure and processes.

Those that we know deviate from such an account will not be useful given the goals and interests of psychology.

An important similarity between both Griffiths' and Machery's position is their insistence that *given the particular aims of the sciences of the mind*, there is a single ideally correct classificatory scheme they ought to adopt. Thus the value of a given psychological or mental category is based on whether it helps us attain the particular investigative goals and aims that are constitutive of the sciences of the mind. But why assume that the sciences of the mind are unified in terms of the properties they investigate, or the goals and aims they pursue? While Griffiths seems happy to grant that there could be a plurality of correct categorizational schemes corresponding to different scientific domains given their different aims and interests, he fails to acknowledge that we find just as much diversity in aims and interests *within* a given scientific field.

With this in mind I propose that, contra Machery and Griffiths, there is no unified set of goals, aims, or methodologies (beyond, perhaps, a very broad and vague commitment to understanding the mind) that dictate a single best categorizational scheme for neuroscience and psychology. While an extremely accurate description of cognitive structures and operations, as well as the identification of natural kinds, might be things that scientists care about, it is clearly not the *only* things they care about, and thus classificatory schemes which fail to fit with these goals will still prove essential to the scientific study of the mind. To this effect, I will argue in the next section for the general idea that different categorizational schemes within the same scientific domain are often essential for learning about different facets of the same complex phenomenon. As such, insisting that a given scientific domain must adhere to a single conceptual framework in order to be of use is impractical and counterproductive.

It is important to note that my intention is not to defend the continued use of the categories ‘emotion’ or ‘concept’ within psychology or neuroscience. Both Machery and Griffiths may be correct that future scientific practice will eliminate both concepts from our lexicon. My intention instead is only to highlight the point that such an elimination cannot be justified on the grounds that these concepts fail to adhere to some ideally correct categorizational scheme regarding the underlying structures and processes of the mind.

2. Plurality Within Scientific Domains

The first thing to note is within any given scientific domain, there are many different aims and goals that scientists have. Scientific interests within a given domain range from hypothesis testing, to prediction, to explanation, to confirmation, to design, to generalization, to understanding, to diagnosis, to the identification of patterns, to pedagogy, to a basis for policy or action (see: Wilson 2006; Giere 2006; Potochnik unpublished: 11). Moreover, we have little reason to assume that the categorizational scheme useful for one of these interests will be useful for others. The best classificatory scheme for the purposes of prediction, for instance, will not necessarily be the one best suited for explanation. Likewise, forming generalizations about the behavior of one system in different contexts may require a different categorizational scheme than one needed for forming generalizations about different systems that display similar behavioural patterns in the same contexts.

To further emphasize this point, consider the aims and goals of physicists when studying the behaviour of fluids. There are different ways of categorizing fluids in the context of physics, and the classificatory scheme physicists use will depend greatly on the particular goals they have in mind when describing the system. As Ronald Giere notes:

> If one is studying diffusion or Brownian motion, one adopts a molecular perspective in which water is regarded as a collection of particles. […] However, if one's concern is the behavior of water flowing through pipes, the best-fitting models are generated within a perspective that models water as a continuous fluid. Thus, one's theoretical perspective on the nature of water depends on the kind of problem one faces. Employing a plurality of perspectives has a solid *pragmatic* justification. There are different problems to be solved, and neither perspective by itself provides adequate resources for solving all the problems. (2006: 34)

Do we classify water as an indivisible continuous fluid, or a collection of discrete interacting particles? Which of these categorizational schemes is most useful for the purposes and aims of physics? It depends on which purpose or aim we have in mind. Do we care about studying diffusion, or water flow? Brownian motion, or wave propagation? We may have good reasons for thinking that, *ontologically speaking*, water is in fact a collection of discrete interacting particles. Thus, we may think that the molecular model more accurately describes the structures and processes underlying the system's behaviour. However, to infer from this that such a categorization of the system would therefore be useful for any and all purposes that physicists might have simply does not follow. Classifying water as a single continuous fluid, one which cannot be decomposed into a collection of discrete particles, is essential if our goal is the prediction of water as it flows through pipes, or if we wish to model the way in which waves propagate when we throw a rock into a pond (Granger 1995: 17; Teller 2001; Thomson-Jones 2005). The scientific value of the classificatory scheme that treats water as a continuous indivisible fluid simply does not depend on whether it properly denotes the underlying natural kinds, or gets the structure of the system right. Since those aren't our concern when trying to predict water-flow, we need not be shackled to the pragmatic and methodological constraints that such a classificatory scheme imposes on us. To insist that physicists should no longer be *permitted* to use continuous-flow models *because* they do not adequately characterize the natural kinds that make up the system, or correctly characterize its underlying causal structure, would be counterproductive to actual scientific practice. By changing the classificatory schemes we use, we can gain different pragmatic benefits useful for the different aims that physicists might have. These are 'situations where data is shifted from one linguistic format to another so that specific forms of question can be more readily addressed' (Wilson 2006: 416-7).

For another example, let us turn to biology. Consider the way in which nerve cells are categorized in a 1984 neurophysiology textbook (Hille 1984: 15). In order to illustrate the electrical features of the cell in a straightforward way to students, the cell is modeled as an electrical circuit, with things like resistors and batteries in place of biological components. As Maria Trumpler notes regarding these sorts of diagrams:

> While the circuit diagram captures the known electrical properties of the membrane, the student should not expect to find actual resistors and batters as cell components. On the other hand, imagining pores in a thick slab omits essential dynamic features. […] It is reasonable to assume that this type of juxtaposition [between the circuit diagram, and the image of pores in a thick slab] was the best way [that authors] knew to elicit a mental imagine in their students that corresponded at least roughly to their own mental images. (1997: 56)

In this case, categorizing the cell in terms of things like batteries and resistors has a *pedagogical* value to biology that a categorizational scheme which only identifies the structure of the actual biological components of the system would lack. And so if our goal is pedagogy, then the best classificatory scheme for biologists to use will not necessarily be the one best suited for the goal of identifying actual biological components. Thus for us to insist that the biological textbook should stop using categories like 'batteries' and 'resistors' because they fail to fit with the categorizational scheme used to characterize the biological structure of systems would be to miss the point of their application in this context (i.e. to better facilitate understanding of the electrical features of the membrane).

But perhaps the case for the study of the *mind* is importantly different. The fact that certain domains of science require different categorizational schemes for different purposes does not mean that this necessarily applies to the scientific study of the mind in particular. Do we have any reason to think that domains like neuroscience, cognitive science, or psychology, use multiple categorizational schemes in the study of the mind? I propose that we do.

3. The Categories of the Mental

Let us return to Griffiths' argument regarding the elimination of the category of 'emotion' from scientific psychology. In order to make his case, Griffiths draws an analogy between the category of 'emotion' in psychology and 'jade' in geology. According to Griffiths, the category 'jade' is inappropriate for the purposes of geology because it partially refers to two different stones. There is no stone 'jade', there are just two different stones with different molecular structures: jadeite and nephrite. Geology, he claims, should therefore abandon the concept of 'jade' and instead talk only of 'jadeite' and 'nephrite'. He then uses this idea to make a parallel argument for the elimination of the concept of 'emotion' from psychology. Note that this sort of argument presupposes that there is a single correct way of categorizing geological systems which respects the causal and structural differences between the two stones (i.e. one which does not involve partial reference). Categories like 'jade' which do not respect these differences fail to fit with this correct geological scheme and thus will not prove fruitful to scientific practice in geology (and likewise with the concept of 'emotion' in psychology).

Despite Griffiths' protestations, however, partial reference of this sort is actually both extremely common, and extremely important, in science. Take jade for instance. While it is true that the category 'jade' partially refers to two different stones with different molecular structures, it is simply false to infer from this that the category of 'jade' therefore has no value in geology. Consider the following passage from a 2005 paper in the *International Geological Review*:

> The term “Jade,” as used in geology and gemology, refers to two extremely tough, essentially monomineralic rocks that are used for fashioning ornamental carvings and gems. Amphibole jade is nephrite, a tremolite-actinolite [$Ca_2(Mg,Fe)_5Si_8O_{22}(OH)_2$] rock with a felted, microcrystalline habit, and pyroxene jade is a jadeite [$NaAlSi_2O_6$] rock (jadeitite) which varies from micro- to macro-crystalline textures. Although their compositions and textures differ, both types of jade are typically associated with serpentinite, and both reflect aqueous fluid crystallization and metasomatic processes. (Harlow & Sorensen 2005: 113)

The authors ultimately conclude ‘not only that the two types of jade share some common petrogenetic characteristics, but that both are products of and record important Earth processes at convergent margins’ (*ibid*: 113-4). In other words, depending on the sort of geological phenomena or properties being studied, the category of ‘jade’ is indeed useful to geologists in virtue of identifying a class of stones which share certain key petrogenetic characteristics as well as certain causal antecedents. Thus the question of whether geologists should, or should not, use the category ‘jade’ depends largely on the interests and goals of individual geologists. Sometimes it is beneficial to think of jade as a single category which includes both jadeite and nephrite given their shared petrogenetic properties. Other times, it is useful to emphasize the differences between the two stones, and thus we are better off adopting a categorizational scheme that does not include ‘jade’, and instead only ‘nephrite’ and ‘jadeite’. To insist that *all* geologists ought to banish the category of ‘jade’ from their scientific vocabulary because it does not fit with one particular way of categorizing geological systems ignores the range of purposes and interests that lead scientists to categorize systems in the first place.

The same lesson applies to our categorization of mental phenomena in domains like neuroscience and psychology. In order to see why, consider for a moment that the way in which a mechanistic system like the brain is implemented matters greatly for what it can do, and under what conditions. The way in which a system is physically structured always affects the way it performs. In this respect, there will always be causal and functional differences between systems that are implemented in different ways, since structural differences always translate into performance differences (for details, see: Betchel & Mundale 1999; Eliasmith 2002, 2013; Bechtel 2008; Syropoulos 2008; Craver 2009). Thus due to differences that exist between individual brains, and between brains of different species, many neurological categories that neuroscientists use will, by necessity, partially refer to differently structured brains with different causal properties. Take, for instance, the hippocampus:

> Consider the hippocampus, a brain region thought to house a mechanism involved in encoding declarative and spatial memories. [...] If one focuses on the particular cells and structures in these specimens, their particular locations and shapes, their particular activities, their exact numbers, and so on, then no two hippocampi are identical. Even those in the same head or in the same location moment to moment are different in many respects. If […] we should split kinds when the mechanisms differ in any of the myriad ways that any two mechanisms might differ—in the precise features of their

> component entities, activities, and organization—then there would be as many kinds of declarative memory encoding mechanisms as there are individual hippocampi (Craver 2009: 585-6).

In order to deal with this sort of worry, Griffiths proposes that we can determine rough criteria for a single correct categorizational scheme in neuroscience that applies across individual brains by appealing to projectability. He says:

> For emotion to be a natural kind [in the sense I mean it], it would need to be the case that the psychological states and processes encompassed by the vernacular category of emotion form a category which allows extrapolation of psychological and neuroscientific findings about a sample of emotions to other emotions in a large enough domain of properties and with enough reliability to make emotion comparable to categories in other mature areas of the life sciences, such as biological systematics or the more robust parts of nosology. (2004: 905)

Thus, he can reasonably claim that even though there are structural and causal differences between individual brains, neuroscientists can still successfully draw inferences about neurological mechanisms across individuals and species from a small sample, and thus such mechanisms can still count as instances of a natural kind. The problem with this idea is that whether or not a brain region or neurological mechanism has sufficient properties in common with those in the brains of other individuals, or of other species, to allow for this sort of inference depends largely on how abstract or detailed our descriptions are, and what the interests and goals of individual scientists are when describing them. Consider declarative memory. Does the mechanism in the hippocampus for encoding declarative memory count as a natural kind in neuroscience? Or does talk of 'declarative memory' co-refer to different kinds of mechanisms in different brains? The answer is: it depends. If we care about broad similarities between individuals, or between species, then an abstract characterization of the mechanism will allow us to draw inferences from our sample to the greater population that meets our purposes. Thus it *is* a natural kind. On the other hand, if we are detailed enough in our description of the structural details, the exact resulting behaviours, and the specific causal properties of the mechanism, then we *won't* be able to generalize to other mechanisms in other individuals or species. Thus it is *not* a natural kind. As Craver notes:

> To solve this problem, it will not suffice to demand that the differences in underlying mechanisms must make a causal difference to (or be otherwise explanatorily relevant to) the behavior of a mechanism as a whole because any detectable difference in the underlying mechanism must make some such causal difference. Likewise, one cannot object that the difference made is too small or insignificant, because such judgments depend on our assessment of which differences are too small to be relevant for our interests, not on the objective features of the causal structure of the world alone. (Craver 2009: 586)[1]

[1] Craver's position in this passage could be interpreted as a sort of spiritual precursor to the position I am defending here, as it nicely highlights the fact that different scientific goals influence the way in which scientists interpret and understand biological mechanisms. That being said, Craver's focus in his paper is the metaphysical

Depending on what we are trying to learn about the system, sometimes structural/causal differences between individual brains matters, and so a categorizational scheme that does not generalize across systems will be better for our purposes. Other times, it is not the differences between individual brains we care about, but their similarities. In which case, treating different mechanisms as instances of the same kind will be required for our purposes. A categorizational scheme that is too focused on highlighting the differences between systems can often obscure the course-grained regularities and patterns that differently structured systems can share, and vice versa (see: Batterman 2002; Potochnik 2010). Thus whether or not something counts as a common mechanism across individuals largely depends on how abstractly we wish to describe the system given our aims and goals. As Bechtel and Mundale note:

> For example, one can adopt a relatively coarse grain, equating psychological states over different individuals or across species. If one employs the same grain, though, one will equate activity in brain areas across species, and one-to-one mapping is preserved (though perhaps further taxonomic refinement and/or delineation may be required). Conversely, one can adopt a very fine grain, and different psychological states between individuals, or even in the same individual over time. If one similarly adopts a fine grain in analyzing the brain, then one is likely to map the psychological differences onto brain differences, and brain differences onto psychological differences. (1999: 202)

nature of mechanistic natural kinds, and not with the question of whether there can be a single ideally correct classificatory scheme for the purposes of any one scientific domain. In this regard, Craver remains relatively silent in his paper on the more general issue being discussed here. It is unclear whether Craver would wish to endorse the position being defended in this paper, but it is certainly consistent with his claims.

Are we trying to see similarities in the behavior and function of different brains and/or highlight broad patterns of behavior across species? Then a broad category will be useful precisely *because* it focuses on similarities and ignores structural and causal differences between the individual brains. On the other hand, if we are interested in an extremely detailed account of how individual brains differ, then a more fine-grained category which does not lump the different neurological mechanisms under the same category will be more suitable. Thus each categorizational scheme will have its place within appropriate scientific practice even though they disagree about whether the identical mechanism counts as a natural kind. Put another way, to insist that a neurological mechanism either *is,* or *is not,* a natural kind for the purposes of psychology and neuroscience implies that there is a single ideal level of abstraction that a correct categorizational scheme should operate on for all neuroscientific and psychological purposes. Yet, what counts as the 'right level' of abstraction depends on the different aims and goals of particular investigators (see, for example: Eliasmith & Trujillo 2014).

So is there one type of mechanism for the encoding of declarative memory, or many different types? If what we care about is the most accurate description of the system's underlying structure and operations, then the best categorizational scheme to use will be a more fine-grained one that does not ignore the structural and causal differences between individual brains. Thus 'declarative memory' will co-refer to different mechanisms. Now according to Griffiths, 'it follows from this fact [that "Jade" co-refers to two different stones] that for the purposes of geology or chemistry, jade cannot be treated as a single kind of thing' (2004: 901-2). If this is true, then for the purposes of neuroscience and psychology, 'declarative memory’ likewise cannot be treated as a single type of mental process that humans share. Does this imply that we should eliminate ‘declarative memory’ from the lexicon of all neuroscience and psychology? Not at all; it just means that the category belongs to a different categorizational scheme useful for different sorts of neuroscientific and psychological purposes (e.g. the identification of more course-grained behavioural patterns displayed by various systems). There are a *plurality* of purposes and aims in neuroscience and psychology, and many require that we classify the system in different ways. And so we cannot simply dismiss a mental category as unhelpful to all of neuroscience or psychology in virtue of failing to adhere to single categorizational framework.

Next, consider Machery’s insistence that any framework that does not accurately describe the natural kinds of a system ought to be eliminated in favour of one that does. He characterizes a natural kind as a class about which many generalizations can be formed because they share at least one common mechanism. Moreover, Machery tells us that a classificatory scheme that does not describe the natural kinds stands in the way of scientific progress by ‘preventing the development of a more adequate classificatory scheme that would identify the relevant natural kinds’ (2009: 230). Yet, as we've seen, whether the mechanisms for declarative memory in different hippocampi counts as a common mechanism or different mechanisms depends on the particular interests and goals of the scientists describing them. So where does this leave neuroscience? Do we eliminate ‘declarative memory’ and ‘hippocampus’ from our neuroscientific vocabulary for being part of a categorizational scheme that *ignores* structural and causal differences between individual brains, or do we instead eliminate the more fine-grained categorizational scheme that *emphasizes* the structural and causal differences between individual brains? Which of these two is the scheme that stands in the way of scientific progress by preventing the development of the more adequate classificatory scheme? To insist that only *one* of these classificatory schemes is allowed to be used, while the other must be discarded, is simply not how actual science works. Neuroscientists use *both* classificatory schemes in their study of the mind. Sometimes the differences between brains *do* matter to the phenomena they wish to study. Other times, they do not. As such, neuroscientists switch between the two types of schemes depending on their purposes and goals. Neither scheme is eliminated in favour of the other.

So should 'emotion' and 'concept' be eliminated from scientific psychology, as Griffiths and Machery suggest? I have no definitive answer to this question, but do hope to have provided some guidance regarding inappropriate criteria for making this decision. Griffiths has argued that the concept of 'emotion' does not fit with the categorizational scheme best suited for the purposes of psychology and the neurosciences. But *which* purposes exactly? John Doris (2000), for instance, argues that the category of 'emotion' may prove essential for different aims and goals in psychology than merely describing the cognitive structures and processes of the brain in detail. Likewise, Lisa Barrett notes that treating emotion as a natural kind for the purposes of classification in psychology and neuroscience may yield tremendous practical benefits, regardless of whether or not it turns out to be a case of partial reference. She notes:

> There are several arguments against dispensing with the natural-kind view of emotion, however. First, this view has been valuable. It is simple to state—emotions are packets of responses that result from mechanisms in the brain and body that derive from our

> animal past—and it is this simplicity that has led to elegant and clear hypotheses that have guided emotion research for almost a century. In fact, the view that emotion categories carve nature at its joints has inspired the research that has produced much of the evidence that we now have, and that we currently argue about (cf. Ekman, 1992). It has also allowed us to make progress on some circumscribed questions (e.g., understanding the neural module for specific behavioral stances in rodents and humans). (Barrett 2006: 46)

Whether such arguments ultimately prove true is a matter of debate, but it highlights the fact that the value of scientific categories is often not determined by whether they do, or do not, fit with some ideally correct categorizational framework.[2]

Conclusion

[2] One might worry that I have defended a position here that is ultimately too permissive, and will let categories into science that might well be worth eliminating. One could, for instance, always jury-rig or contrive a scientific context in which it might be pragmatically useful to employ any number of unscientific categories or concepts, such as 'phlogiston', 'entelechy', or 'witches' (in order to make limited predicts, for instance, or motivate certain behaviours). It seems that any number of categories that science has justifiably eliminated can be salvaged in this manner.

But this objection would be to misunderstand my argument. I am not claiming that any category which can be shown to have limited pragmatic value under specified circumstances is thereby indispensable to science. In fact, I am making no claims as to the criteria needed for scientific indispensability (as I noted above, it might well prove to be the case that categories such as "emotion" and "concept" are worthy of elimination).

My point instead is to deny that a justifiable or appropriate *criterion* for scientific elimination is whether a category fails to cohere with some ideally correct categorizational scheme. I have argued that the plurality of goals that scientists have necessitates the application of distinct classificatory schemes, and that there is no one categorizational scheme useful for every goal that scientists have in a given domain. Thus we cannot justify the elimination of a given category based merely on the fact that it happens not to cohere with one particular way of categorizing phenomena. This fact, however, does not imply that any category is as good as any other for our scientific purposes.

It is commonly held in the philosophy of mind and philosophy of psychology that there is a single best categorizational scheme for the purposes of psychology and neuroscience. If a mental category fails to adhere to this ideally correct scheme, then it should be eliminated from scientific practice. In this paper, I have argued that this type of view is misguided and runs counter to scientific practice. Scientists often employ a plurality of categorizational schemes depending on their needs and goals in a given context. There are features of complex systems that we simply cannot learn by employing only a single categorizational scheme, and so a plurality of schemes provides a range of pragmatic benefits. To suggest that a given mental category ought to be abandoned because it does not have a place in a particular categorizational scheme undermines a good deal of productive scientific practice in neuroscience and psychology.

Acknowledgements: I would like to thank Beate Krickel, Anya Plutynski, Carl Craver, Carla Fehr, Doreen Fraser, and Chris Eliasmith for helpful discussions and feedback on earlier versions of this paper. A much earlier draft was also presented at *iCog*: *Interdisciplinarity Conference in Cognitive Science* under a different title in December 2013. Feedback from the audience at this conference was extremely helpful in shaping this paper. This research was supported by the Social Sciences and Humanities Research Council of Canada Postdoctoral Award.

References

Barrett, L. (2006) ‘Are Emotions Natural Kinds?’, *Perspectives on Psychological Science,* 1/1: 28-58.

Batterman, R. (2002) ‘Asymptotics and the Role of Minimal Models’, *British Journal for the Philosophy of Science,* 53: 21-38.

Bechtel, W. (2008) *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. New York: Lawrence Erlbaum Associates.

Bechtel, W. and J. Mundale (1999) ‘Multiple Realizability Revisited: Linking Cognitive and Neural States’, *Philosophy of Science,* 66/2: 175-207.

Bickle, J. (1998) *Psychoneural Reduction: The New Wave*. Cambridge, MA: MIT Press.

Bickle, J. (2003) *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Boston: Kluwer Academic Publishers.

Bickle, J. (2006) ‘Reducing Mind to Molecular Pathways: Explicating the Reductionism Implicit in Current Cellular and Molecular Neuroscience’, *Synthese,* 151: 411-434.

Churchland, P. M. (1981) ‘Eliminative Materialism and the Propositional Attitudes’, *The Journal of Philosophy,* 78: 67-90.

Craver, C. (2009) ‘Mechanisms and Natural Kinds’, *Philosophical Psychology,* 22/5: 575-594.

Devitt, M. and K. Sterelny (1987) *Language and Reality: An Introduction to the Philosophy of Language.* Cambridge, MA: MIT Press.

Doris, J. (2000) ‘Review of Griffiths, Paul E. What Emotions Really Are: the problem of psychological categories’, *Ethics,* 10/3: 617-619.

Eliasmith, C. (2002) ‘The Myth of the Turing Machine: The Failing of Functionalism and Related Theses’, *Journal of Experimental & Theoretical Artificial Intelligence,* 14/1: 1-8.

Eliasmith, C. (2013) *How to build a brain: A neural architecture for biological cognition*. Oxford University Press.

Eliasmith, C., and O. Trujillo (2014) 'The use and abuse of large-scale brain models', *Current Opinion in Neurobiology*, 25: 1–6.

Giere, R. (2006) 'Perspectival Pluralism', in S. Kellert, H. Longino, and C.K. Waters (eds.) *Scientific Pluralism*, 167-190. Minneapolis: University of Minnesota Press.

Granger, R. A. (1995) *Fluid Mechanics*. New York: Dover.

Griffiths, P. (1997) *What Emotions Really Are: The Problem of Psychological Categories*. Chicago: University of Chicago Press.

Griffiths, P. (2004) 'Emotions as Natural and Normative Kinds', *Philosophy of Science,* 71/5: 901-911.

Harlow, G.E., and S.S. Sorensen (2005) 'Jade (Nephrite and Jadeitite) and Serpentinite: Metasomatic Connections', *International Geology Review,* 47: 113-146.

Hille, B. (1984) *Ionic Channels of Excitable Membranes.* Sutherland, Mass: Sinauer.

Levy, N. (2011) 'Resisting 'Weakness of the Will'', *Philosophy and Phenomenological Research,* 82/1: 134-155.

Machery, E. (2009) *Doing Without Concepts*. Oxford: Oxford University Press.

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

Penn, D. (2012). 'How Folk Psychology Ruined Comparative Psychology: And How Scrub Jays Can Save It', in R. Menzel and J. Fischer (eds.) *Animal Thinking: Contemporary Issues in Comparative Cognition*, 253–266. Cambridge: MIT Press.

Potochnik, A. (unpublished manuscript) *Idealization and the Aims of Science*.

Potochnik, A. (2010) ‘Explanatory Independence and Epistemic Interdependence: A Case Study of the Optimality Approach’, *The British Journal for the Philosophy of Science,* 61/1: 213-233.

Ramsey, W., Stich, S. and J. Garon (1990) ‘Connectionism, Eliminativism, and the Future of Folk Psychology’, *Philosophical Perspectives,* 4: 499–533.

Sehon, S. (1997) ‘Natural-Kind Terms and the Status of Folk Psychology’, *American Philosophical Quarterly,* 34/3: 333-344.

Stich, S. (1983) *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge, Massachusetts: The MIT Press.

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

Syropoulos, A. (2008) *Hypercomputation: Computing Beyond the Church-Turing Barrier*. Springer.

Teller, P. (2001) ‘Twilight of the Perfect Model Model’, *Erkenntnis,* 55**:** 393–415.

Thomson-Jones, M. (2005) ‘Idealization and Abstraction: A Framework’, *Poznan Studies in the Philosophy of the Sciences and the Humanities,* 86/1: 173-218.

Trumpler, M. (1997) ‘Techniques of Intervention and Forms of Representation of Sodium-Channel Proteins in Nerve Cell Membranes.’, *Journal of History of Biology,* 30/1: 55-89.

Wilson, M. (2006) *Wandering Significance: An Essay on Conceptual Behaviour.* New York: Claren.