Shame and Philosophy

An Investigation in the Philosophy of Emotions and Ethics

Phil Hutchinson
Manchester Metropolitan University

palgrave
macmillan

Contents

Introduction

Ethical enquiry takes a number of forms. It can be conducted in the manner of normative moral theorising: theorising as to how we ought to act and/or live. It can be conducted in the manner of metaethical enquiry: enquiring about the nature of value and value claims. It can be conducted in another, more tangential, though, I suggest, equally as important a manner; not directly concerned to theorise as to what one ought to do, nor concerned directly with the metaphysics of value, but rather concerned with questions such as what it means to be human, what place do moral concepts have in our lives, and how are they related to other concepts. This is the sense in which the present work is an ethical investigation; it is offered as a work in moral psychology, though, importantly, one which also seeks to work on the reader's moral sensibilities.[1]

The main title of the book has a double meaning, as no doubt will have been suspected. The first of these is the 'straight', or literal, meaning; here the topic of investigation is 'shame', the emotion, and the investigation is philosophical. The second meaning, though less literal, is no less central to my purpose; here I am suggesting for discussion that philosophy, as a subject, should feel a little ashamed. This, of course, is a deliberately dissentious claim. I mean to suggest that many of us, when doing philosophy, continually acquiesce to the temptation to abstract from personhood, from the world we inhabit, etc. in order to reflect on matters which have direct impact upon, and are directly related to, the person.

Recently there has been a revival of interest in the virtues. We are often told that we now have three approaches to normative ethics from which to choose.[2] This might well be so. What is appealing about virtue ethics is the central place such an approach accords to personhood, character,

and the life lived with others. This is the topic, somewhat tangentially addressed, of the present work. It is hoped that the reflection on the emotions, on shame, and on philosophers who have undertook to study them is one that will give rise to one's reflection on the nature of, and conditions for, being a person. One way in which I hope it will be so is by, as the work progresses, increasingly engaging with the emotion of shame as experienced, documented, and reflected upon by survivors of (and one perpetrator of) extreme trauma. This serves to concretise the discussion.

As regards shame, my aim is not to provide a theory of shame but rather a framework for understanding. My philosophical approach is therapeutic; in this regard my objective is purely to facilitate understanding. I will not theorise the substantive content of shame (or the emotions in general). Such content comes from our observation of the phenomena, clearly viewed. The philosophical task I undertake is to provide a framework that facilitates our understanding of the phenomena. I seek no more than to aid the reader to see shame by providing a perspicuous presentation of the phenomena in question; this is what my framework for understanding—the 'world-taking cognitivist' approach—seeks to make possible.

I will not, therefore, be concerned to arrive at a conclusion that offers a substantive theory of the emotions, of shame, or of the person. Indeed, I shall offer no theory of the person. What I seek to show is that shame as an emotion, when understood, can afford us insight into our nature as human animals, our nature as persons. One of the promises of virtue ethics is that the person is 'brought back in'; unfortunately, the promise is not always kept, because in seeking to compete with the other two normative theories, the virtue ethics on show often becomes virtue *theory*, merely one methodological tool among others that can be applied by the theorist faced with a moral dilemma. In what follows, I propose to meditate upon both personhood *and* philosophy's relationship to personhood. I do this through an analysis of the emotions and of shame. 'Signposting' and the explicit stating of aims and objectives will be kept to a minimum; the therapeutic approach to philosophy seeks to facilitate aspect shifts—reorientations in thought—in one's readers. Such reorientations must be *freely* arrived at. The author's—my—task is, therefore, one of *facilitating* such aspect shifts.[3]

This book has a structure that needs some explanation. Chapters 1 and 2 engage with philosophical approaches to emotion, which might well be considered to be at polar extremes. While Chapter 3 works towards a way of understanding emotion through an engagement with,

what are most often termed as, cognitive theories of emotion. Chapter 4 begins by addressing some of the current issues in philosophical discussions of the emotions and moves towards conclusion by suggesting future directions of study.

The book will take the form described below.

Chapter 1

Experimental methods and conceptual confusion: philosophy, science, and *What Emotions Really Are*

My reflections on emotions begin with an engagement with a hugely influential book: Paul E. Griffiths' *What Emotions Really Are*. This is done, and takes such a place of prominence, for a number of reasons. Griffiths, in effect, denies philosophers a voice when considering emotion. He seeks to show that philosophical theories of emotion are nothing more than recapitulations of the current stereotype of emotion terms, sometimes stated as being nothing more than recapitulations of our folk psychology of emotions. One needs to meet Griffiths' challenge. Griffiths' book is a prominent, influential, and sophisticated version of philosophical scientism. It can be tempting to see scientism as self-loathing philosophy—philosophy that cannot tolerate acknowledgement of anything genuinely and distinctively philosophical, i.e. philosophy that is not modelled on or reducible to (what are taken to be) the methods of or the results of natural science. Griffiths both denies that (non-scientistic) philosophy can give any insight into the reality of emotion and advocates a science of emotion.

An engagement with Griffiths' work is a pertinent place to begin. I give some time to outlining Griffiths' claims; I then set about identifying the presuppositions which lead him to make those claims. Ultimately, there is in play an eliding of normativity. This leads to a replacement of ordinary talk of 'learning' with talk of 'phenotypes developing'; it leads to a replacement of talk of the place of emotions in people's lives with talk of 'affect programs running on a limited database'. Now, it is easy and may be tempting to sneer at the employment of such language if one is not tempted by scientism; the burden of Chapter 1 is to resist such sneering and rather try to bring us to the realisation that such language plays no more than a rhetorical role. The claim is *not* that authors such as Griffiths, in a somewhat Machiavellian manner, deliberately set out to blind us with rhetoric; rather, the claim is that if we can show such use of language to have no more than rhetorical significance, over and above

that which it replaces (and often less significance in terms of making sense of our lives), we might persuade such authors and those persuaded by them to rethink their position.

Griffiths' scientism is founded upon a scientifically determined theory of meaning: his own version of the causal homeostatic theory of natural kind semantics. The chapter ends by identifying the limits of such a theory of language and with a suggestion about concepts, and how we might understand their place in our lives.

Chapter 2

To 'Make Our Voices Resonate' or 'To Be Silent': Shame as Fundamental Ontology

Having engaged with scientism in the philosophy of the emotions, one might recoil. Such recoil might lead one to an engagement with a radically different form of philosophy, a form of philosophy very distant from scientism. Such a form of philosophising might both reject (what are understood as) the methods of the natural sciences as having import for philosophical reflection and might eschew attempts at determining meaning. Giorgio Agamben's work provides the locus for an engagement with such a philosophical position. Agamben situates his work in the post-Heideggerian tradition, and his discussion of shame is embedded in a wider historico-political thesis. His claim is that the individual's awareness of the self is felt as shame. In making this claim he draws upon Aristotle, Derrida, Foucault, Heidegger, Kant, and Levinas and claims that phenomenological support is provided by an analysis of Holocaust survivor testimony, particularly that of Primo Levi and of Robert Antelme.

Such a project initially shows promise, for it is somewhat refreshing to be moved away from the abstractions of scientism. However, such promise is not realised in the final analysis. Levi and Antelme are both garnered as support; unfortunately their own writings do not quite offer the support that Agamben presents them as so doing. Noting this serves as a spring-board to reflection upon Agamben's own philosophical prejudice. While engaging in a mode of philosophising which one might see as standing diametrically opposed to scientism, one finds it to be a mode of philosophising which can similarly elide the person in the name of theory. Ultimately this eliding is driven by a picture of language, which too-readily abstracts from the person's role in language. The eliding is, therefore, structurally conferred.

Agamben's prejudice is, we might say, to be located in his indebtedness to post-structural linguistics. Here, in contrast to the attempts to theoretically determine meaning that we found in scientism, meaning is said to be continually deferred, always—logically—just beyond our grasp.

The philosophical task is to make emotional expression intelligible, to elucidate its place in our lives. Agamben's inquiries do not fulfil this task. The cost of this failure is misrepresentation of expressions of shame.

Chapter 3

Emotion, Cognition, and World

Chapters 1 and 2, in engaging with visions of the philosophical task which appear to be polar opposites regarding the understanding of our subject and approach to the subject matter, help to orient me as regards my own enquiries. I want to understand shame as it is expressed by Primo Levi, and other survivors of extreme trauma; to gain some understanding of the place of shame in the lives of people; and to gain some understanding of the nature of the person. Neither the approach recommended by Griffiths nor that advanced by Agamben is sufficient to this task (though one gains much more from the latter of the two). However, in recognising their deficiencies, I am better placed with regard to pursuit of my goal.

Chapter 3 pursues this goal. In Chapter 1, I note how Griffiths focuses his hostility to philosophical accounts of the emotions on an approach to emotions called cognitivism. In Chapter 3, I discuss variants of cognitivism. In the course of doing so, I explore numerous ways in which philosophers of the emotions marginalise the person, through the invocation of sub-personal mechanisms. I suggest that such a denial is borne of having in play a picture of mind and world which is *not*, on reflection, obligatory. This picture underlies traditional cognitivism (what I term 'reason-giving cognitivism'), and it underlies many of the critical remarks offered by critics of traditional cognitivism.

Chapter 3, therefore, sees me pursuing—broadly speaking—two tasks: first I seek to make manifest the pictures, the thought-constraining grip of which leads both reason-giving cognitivists and their opponents to their conclusions; second I seek to offer another way of understanding emotion: (what I term) 'world-taking cognitivism'. World-taking cognitivism, I suggest, avoids the problems to which the other philosophical accounts of emotion are subject. Most importantly, it is a way of understanding emotion in which the person is central, not marginalised by a desire for non-normative explanation (Griffiths; Chapter 1),

not marginalised by the search for a fundamental ontology (Agamben; Chapter 2), and not marginalised by the invocation of sub-personal mechanisms (Robinson, Prinz; Chapter3).

Chapter 4

Shame and World

Chapter 4 moves us towards conclusion. I first subject to scrutiny some criticisms of cognitivism not covered in Chapter 3. This is the criticism advanced by John Deigh (and many others) that a commitment to the fact of the intentionality of emotion cannot be aligned with the fact of (non-human) animal emotions. I show that Deigh's 'two facts' do not pose a problem for the approach to emotion I offer here.

I then progress to a more focussed discussion of shame. I pay particular attention to the metaphor of the audience and to the question of heteronomy, as discussed by Bernard Williams (1993). I move towards the concluding section with a discussion of bystanders and the absence of shame.

It is my hope that by this stage the ethical sense of this work, which I discussed above, will have become manifest. It is ethical in the sense of my remarks about the second way in which one might understand the title. Philosophies which proceed to offer explanations which elide the person are all too familiar; as a philosopher one should make every attempt to find a way of describing a situation in a perspicuous manner, before turning to theory. It is also ethical in the sense in which it provides, I hope, a small step on the path to a better understanding of the person and, through the examples garnered, stimulates one's moral sensibilities.

1

Experimental Methods and Conceptual Confusion: Philosophy, Science, and *What Emotions **Really** Are*

Philosophers in the Anglo-American tradition began to take renewed interest in the emotions in the 1960s. Since then the dominant 'research program' in the philosophy of the emotions has been—what is widely, though not uncontroversially, called—cognitivism. Authors such as Anthony Kenny (1963), Robert Solomon (1976, 2003c), Gabriele Taylor (1985), and Peter Goldie (2000)[1] have offered explanations of the human emotions chiefly in terms of the beliefs (thoughts, judgements, evaluations) of the agents; in the early stages this 'project' was seen (often self-consciously) as a corrective to 'feeling theories' of the emotions, particularly those offered and/or influenced by William James (1884) and Carl Lange (1885)—often referred to as the James-Lange theory—which depicted emotions in a manner which led to them being characterised as irrational irruptions into an otherwise rational life. Cognitivism was seen as a corrective to this, in that it set out to rationally explain the emotions. Recently, the post-1960s, dominance of philosophical cognitivism has been subjected to strong criticism.

In his (1997) book *What Emotions Really Are* (*WERA*), Paul Griffiths launches a 'blistering' attack on philosophical cognitivism in the philosophy of emotions.[2] Griffiths refers to cognitivism in the philosophy of the emotions as 'propositional attitude theory' so as to guard against it being confused with cognitive science; the latter being an area of inquiry which he thinks has genuine merit. In this chapter I will follow Griffiths' terminology, if only so as to avoid confusion. (I shall critically examine his choice of the name 'propositional attitude theory' when I examine 'cognitivist' accounts in Chapter 3).

I examine Paul Griffiths' work for a number of (related) reasons: first, because it provides us with a particularly stark and very forthright example of scientism in the philosophy of the emotions, and philosophy

in general; second, Griffiths' book has been extraordinarily influential;[3] and third, if Griffiths is correct most of the philosophical work on the emotions undertaken over the last 30 years has told us nothing about what emotions really are. As with regards to 'scientism', this is a term often employed pejoratively; this is not my intention here. Every aspect of Paul Griffiths' book is unabashedly, and somewhat defiantly, scientistic.[4] 'What emotions really are' will be explained by a science of the emotions. Griffiths' book is part (Part 1 of the book) report on the current state of the 'science' of emotions and part (Part 2 of the book) argument for a theory of language which will allow for (improved prospects for) scientific explanation of the emotions.

'What emotions really are', for Griffiths, is not what they mean for the individuals experiencing them; not what an examination of the use of emotion terms might tell us; they are not explained by some combination of the propositional beliefs and the desires of the individual experiencing the emotion; nor are they explained by some combination of the propositional beliefs, the desires, and the attendant feelings of the individual experiencing the emotion. Indeed, neither does it help—as far as Griffiths is concerned—if we include a bit of added narrative; nor if we replace 'beliefs' with 'judgements', 'evaluations', or 'construals'. No; 'what emotions *really* are' is explained by uncovering and scientifically explaining the category to which the concept (putatively) refers.

Griffiths makes many (substantive) claims in his book. I focus my attention on his main philosophical claim. Both his criticisms of propositional attitude theory and his proposals for the future scientific study and explanation of emotion explicitly rest upon his arguments and proposals for a causal homeostatic theory of natural kind semantics. Griffiths' work in the emotions, therefore, stands or falls on the arguments he advances in Chapters 7 and 8 of *WERA*.

This chapter takes the following form:

- Section 1 gives a *précis* of *What Emotions Really Are*. In this section I restrict critical comments of my own to a minimum. It is tempting to engage with the substantive claims Griffiths makes both about the emotions, in Chapters 2 through 6 (Part 1) of his book, and about progress in philosophy throughout his book; however, the temptation is (in the main) resisted. The force behind Griffiths' criticisms of propositional attitude theory and his substantive claims about what emotions *really* are, as noted, arises from his theory of natural kinds.
- Section 2 critically engages the history of natural kind semantics, as that history is presented to us by Griffiths. I begin by looking at Hilary

Putnam's (1975d) account: Griffiths claims to be working within the paradigm initiated by Putnam (and Kripke 1980 [1972]). I question Griffiths' rendition of Putnam; I advance some of my own criticisms of Putnam's position (correctly understood); and I note the criticisms of Putnam advanced by others.

- Section 3 critically engages Griffiths' own theory of natural kinds: the causal homeostatic theory. Griffiths claims his theory avoids the pitfalls to which *he* (see Section 2) found Putnam's to be subject. I question this. Griffiths oscillates between, on the one hand advancing a robustly metaphysical thesis,[5] and on the other hand doing no more than describing the pragmatic nature of our concepts in certain domains.[6]
- Section 4 moves towards conclusion by suggesting that careful reflection upon our concepts not only gives answers to questions, which Griffiths *assumes* only science can answer, but also shows us the conceptual confusions in which Griffiths seems trapped.

1. Précis of *What Emotions Really Are*

1.1. Degenerative research programs

Griffiths' charge is substantial. Not only does he see in propositional attitude theory a resounding and inevitable failure, but also charges its proponents with a systematic antipathy for the results of scientific psychology. The antipathy is 'systematic' because the failure to acknowledge the 'insights' of 'science' is no mere accident or oversight on the part of propositional attitude theorists; rather, it is a result of their methodological orientation. While propositional attitude theory in the philosophy of the emotions has its origins in Anthony Kenny's (1963) *Action, Emotion and Will*, it had still, 27 years later (in 1989), failed to overcome its original difficulties. The conclusion should be that propositional attitude theory is a failed research program (Griffiths 1989; *WERA*: p. 38).

Griffiths' attack is two-pronged. The first prong, we might say, is his identification of a number of substantive failures of propositional attitude theory. These failures are: the inability to account for objectless emotions, reflex emotions, and unemotional evaluations; further charging that the theory cannot provide an explanation for the underdetermination of emotions by judgements, emotional responses to imagination, and physiological responses. Despite some late attempts by some propositional attitude theorists (e.g. Stocker 1987; Nash 1989) to address some of these problems, Griffiths suggests that these substantive problems cannot be overcome by the research program for deep-seated,

methodological reasons. This is the second prong of Griffiths' attack. For, one might be tempted to defend the propositional attitude theorists by asking for them to be granted more time in which they might address these substantive problems. However, such a (potential) plea is undercut by the charge that these substantive failures stem from an underlying, fundamental, methodological flaw: i.e. the theory's reliance upon conceptual analysis.

Conceptual analysis as a methodology comes in for a similar two-pronged attack. The first prong is Griffiths' *assertion* that the form of conceptual analysis engaged in by propositional attitude theorists 'is a view which has been very broadly rejected in the philosophy of language' (*WERA*: p. 4); the second prong is the charge that there are flaws in the account of linguistic meaning presupposed by conceptual analysis. Conceptual analysis, we are told, presupposes that a concept can be defined by identifying its rules for correct application; therefore, it should be rejected as a methodology owing to its failure to do more than tell us the current stereotype of a concept. Equating the meaning of a concept with the current stereotype leads to our explanation of (say) 'fear' being no more than a reflection of the current stereotype of fear. The (alleged) problem with this is that stereotypes change as our knowledge of things (or phenomena) grows. Hence, allowing one's explanation of a phenomenon to rest upon the stereotypical understanding of that phenomenon is merely to explain the phenomenon in terms of our current (contingent) understanding of it, and not to explain the phenomenon itself. This is, we are told, as if to define a whale as a fish because people (folk) once thought that whales were fish. Griffiths writes:

> All conceptual analysis will reveal is the current stereotype of fear. To insist that all and only the things that fit this stereotype are examples of the kind is simply to stand in the way of clarifying the concept. It is exactly akin to insisting that whales are fish because people called them so.
>
> (*WERA*: p. 5)

In place of conceptual analysis Griffiths proposes the semantics of natural kinds, conceived by Kripke (1980) and Putnam (1975d), but developed and refined since.[7] Briefly (I come back to this in more detail below), the semantics of natural kinds classifies a term into four components: syntactic marker, semantic marker, stereotype, and extension. An example of these classifications for the term 'whale' might be as follows: syntactic

marker—'noun'; semantic marker—'mammal'; stereotype—(something like) 'large, migratory, sea-faring mammal'; extension—what the best current science tells us a whale 'really is'. On Griffiths' understanding we can get by in our day-to-day lives, and be seen as competent users of the language, while only ever knowing the stereotype of a term. However, to really know the *meaning* of a term is to know its extension, and the knowledge of a term's extension is identified with the best current scientific knowledge of the kind. I shall leave to one side, for now, what 'scientific knowledge' might mean here, but, needless to say, what it might mean is central to any account given of the semantics of natural kinds (I shall address this in Section 3.5). For now it is enough to note that for Griffiths conceptual analysis can only alert us to the stereotype of a term; in order for us to know what emotions *really* are we need to know more than the current stereotype.

To recap, Griffiths sets his sights on the dominant philosophical approach to the emotions: the propositional attitude theory. He depicts a number of theorists, whom he claims advance a propositional attitude theory of the emotions, as comprising a research program. He charges the research program with a number of serious substantive failures to explain the emotions, arguing—as he has before (Griffiths 1989)—that these substantive failures are *in themselves* enough for us to conclude that the research program is no longer worth pursuing (it is a terminally degenerative research program). He then submits the further charge that these substantive failures are borne of the research program's methodological reliance upon conceptual analysis:[8] writing that conceptual analysis has been largely abandoned as a research program in the philosophy of language, epistemology, and the philosophy of mind. If, therefore, we want to answer the 'vernacular' question 'what *are* the emotions?' (ibid.: p. 228) we need to do more than identify the stereotype, we need to know the extension of the term. The subsequent chapters of Part 1 of Griffiths' book explore the ways in which—what he takes to be—current science might 'fill out' the extension.

1.2. Alternative approaches: learning from 'science'

Griffiths identifies three distinct phenomena that fall under the vernacular term 'the emotions': affect program emotions; higher cognitive emotions (which in subsequent publications he prefers to call 'complex emotions'); and disclaimed action 'emotions'.

1.2.1. Affect program emotions

'The *affect program* is the coordinated set of changes that constitute the emotional response' (*WERA*: p. 77) and these changes are experienced as psychological events.[9] The leading exponent of this approach is Paul Ekman (1975),[10] whose work builds upon Charles Darwin's nineteenth-century experiments on facial expression. Ekman (and various co-authors) have conducted experiments on human facial expressions across cultures; from these experiments they hypothesise that certain emotion terms are the names of categories of psychological event. The hypotheses, we are told, are confirmed in ongoing experiments whereby people are asked to examine photographs of a number of facial expressions and attribute the emotion being experienced by the person in the photograph. Ekman and his various co-authors claim that these experiments uncover 'six species-typical' human affect programs: surprise, anger, fear, disgust, sadness, and joy (*WERA*: p. 78). Griffiths claims that Ekman's research is complemented and supported by work done on the autonomic nervous system (ANS) by Joseph Schachter (1957), Gary E. Schwartz *et al*. (1981), and Antonio Damasio (1994); these authors' work shows that the arousal of the ANS is differentiated among the emotions in a way that lends support to the hypotheses advanced in Ekman's research. Those emotions which are explained as affect programs are therefore 'restricted to short-term, stereotyped responses, triggered by modular subsystems operating on a limited database' (*WERA*: p. 241).

1.2.2. The higher cognitive emotions

The higher cognitive emotions are, paradigmatically, emotions such as pride, shame, guilt, and remorse (Griffiths includes loyalty and revenge in the list). These differ from the affect programs in that they are 'irruptive motivations' (*WERA*: p. 243). That is to say, the higher cognitive emotions are irruptions in our long-term planned actions, owing to our immediate circumstances. While Griffiths told us that the affect program emotions are 'sources of motivation not integrated into the system of beliefs and desires' (ibid.) thus requiring the introduction of the concept of *mental state*, the higher cognitive emotions *are* (often) integrated into our beliefs and desires. Griffiths largely[11] follows Frank (1988) in seeing in the higher cognitive emotions *apparently* irrational responses to our immediate environment, which enable us to pursue long-term rational goals. So, for example, we believe we should be loyal, owing to current circumstances; therefore, we depart from our goal-directed long-term plans in order to carry out the duties demanded by that loyalty.[12] Those duties,

though seemingly irrational in virtue of our long-term plans, might serve a rational purpose in engendering loyalty, trust, and the like in others. Thus long-term rational plans are served after all. This, we might note, relies heavily on the 'findings'[13] of game theory. The higher cognitive emotions, then, are less like reflex responses than the affect program emotions.[14] Griffiths sees the scientific explanation of the higher cognitive emotions as being more troublesome. Current work in evolutionary psychology and game theory *might* combine to provide the answers, but many questions are, at present, pending satisfactory answers. What is clear to Griffiths, however, is that we have enough evidence to establish the existence of a category of emotions distinct from the affect program emotions.

1.2.3. Disclaimed action emotions

Disclaimed action is uncovered on examination of the literature on the social construction of emotion[15] (*WERA*: ch. 6). Griffiths divides the social construction of emotion into two models: the social concept model and the social role model. We are told that the social role model has two variants: the disclaimed action version and reinforcement version (*WERA*: p. 143). The disclaimed action version of the social role model is the aspect of social constructionist accounts that Griffiths thinks offers insight into some 'emotion' phenomena; such insight is not afforded by either the affect program theory or the psycho-evolutionary candidate for explaining the higher cognitive emotions. In a disclaimed action, the behaviour of an individual is 'acted' in an attempt to conform to a social role. This differs from the affect program emotions and the higher cognitive emotions in that in the first instance it lacks the professed cross-cultural status of the other two. Examples of such 'culturally specific' disclaimed action are: the state of 'being a wild pig' reported in the Gururumba people of New Guinea (Newman 1964, cited in *WERA*: p. 140); 'running *amok*', which has been documented in South East Asian societies; and 'multiple personality syndrome' (MPS) found in some western societies, as discussed by Ian Hacking (1995).[16] Emotions can be disclaimed actions in the same way that 'being a wild pig', 'running *amok*', and 'multiple personality syndrome' are identified as being. Disclaimed action emotions are a display of behaviour that is socially appropriate in a particular situation. This is learned behaviour, though neither the individual nor society acknowledges this fact. On the contrary, the behaviour is seen as 'a natural and inevitable response to the circumstances and outside the control of the individual' (*WERA*: p. 141).

The disclaimed action emotions are pseudo-emotions on Griffiths' account. They are strategic in that they play on the status accorded to particular emotions in certain circumstances in different societies. If a given situation is acknowledged by society as giving rise to extreme anger, and extreme anger is acknowledged by a society as being worthy of providing mitigation for an act, then we might find anger behaviour acted out within that society in those situations. This anger is not 'real' anger—affect program anger—but rather disclaimed-action-'anger'. Importantly, the suggestion is not that this is a self-conscious, calculated 'aping' of the emotion on the part of the individual, rather it is *caused* by internalised (learnt) beliefs about appropriate action in given situations in certain societies. Griffiths tells us,

> Disclaimed action emotions are modelled on the local cultures' conception of the emotions. They aim to take advantage of the special status that emotions are accorded because of their passivity. Like socially constructed illnesses, disclaimed action emotions are actually very different from the phenomena on which they are modelled. At a psychological level, far from being disruptive of longer term goals, they are 'strategic' devices for the achievement of those goals. Rather than involving isolated modules [affect programs], or special adaptations of higher level cognition [higher cognitive emotions], they are manifestations of the central purpose of higher cognitive activity—the understanding and manipulation of social relations.
>
> (*WERA*: p. 245)

Disclaimed actions are, then, distinct from both affect program emotions and the higher cognitive emotions. What is in question is whether they are to be admitted to the list as a genuine emotion category: i.e. whether they form a natural kind; Griffiths suspects not, though he leaves the question (partially) open.

1.3. Eliminating emotion

Griffiths' discussions in Part 1 of *What Emotions Really Are*, together with his declared philosophical agenda, lead him to conclude that the vernacular term 'the emotions' needs to be eliminated in favour of two natural kind terms corresponding to the two identified categories of emotions. The elimination of the 'folk concept' will better facilitate the induction and explanation of the emotion. We cannot identify the vernacular term 'the emotions' with one of the above kinds of emotional explanation; for if we were to identify the vernacular term 'the emotions'

with the affect program emotions we would be forced to exclude the higher cognitive emotions. This would lead to a failure to answer the question.

It strikes me that there is a degree of confusion in play here. Either the concept of 'emotion' employed in the vernacular question 'what is an emotion?' has significance, or it has not. On the one hand he wishes to say it has not, and should therefore be eliminated—because it does not refer to one natural kind, but two, maybe three; while on the other hand he says we cannot answer the question 'what is an emotion' in a way which leaves out a number of concepts which we would ordinarily take to be emotion concepts—i.e. the paradigm cases of higher cognitive emotions: shame, guilt, remorse, &c.—because in doing so we would have failed to have adequately answered the (vernacular) question. Is Griffiths here—in the latter case—saying it is part of what we mean by 'emotion' that we include 'both' higher cognitive *and* affect program emotions in the extension of the term, and thus we would not admit an answer which excluded 'one' of 'these'? It pays also to give thought to the distinction between the higher cognitive emotions and the, so called, disclaimed action 'emotions'. The distinction rests upon two claims: first, that the disclaimed action emotions are consonant with long-term rational goal-directed actions as opposed to disruptive of them, as in the case of the higher cognitive emotions; and second, the claim that these (latter) 'emotions' cannot be genuine because they are culturally indexed. These are the reasons provided for these 'emotions' forming a distinct 'category'. To refuse to accept them as real emotions for these two reasons is to already have in play an account (a substantive conception) of what counts as *real* hereabouts.

Furthermore, the question 'what are the emotions?' is, Griffiths tells us (p. 242), a request for an answer which distinguishes the emotions from other cognitive processes. Given what we have learnt of his project thus far, we might expect Griffiths to hold that there is no possible answer to such a question. However, he informs us that there is, and it is the oft-cited phenomena referred to as the 'passivity' of emotion. He writes:

> What is to be explained by emotional phenomena in general is the way in which they contrast to other cognitive processes. The phenomena referred to as the 'passivity' of emotion are central to this contrast. I am not convinced that all instances of the passivity phenomena can be explained by the modularity of the affect programs. I suggested in chapter 5 that a form of passivity may characterise some emotional responses controlled by higher cognition. These responses

> are *irruptive* motivations: motivations not derived from more general goals by means-end reasoning. This class of states has as good a claim to be the referent of the general concept of emotion as the class of affect program states.
>
> (ibid.)

The point here is somewhat obscure. On the one hand he—seemingly approvingly—claims that emotions can be explained in terms of their passivity, i.e. it is this that differentiates them from other cognitive processes; [17] yet on the other hand, he says that the modularity of affect programs cannot explain all instances of such passivity. He seems to assume that this inability on the part of computational psychological explanations of affect programs to explain passivity means we should question the explanatory/epistemic worth of passivity, rather than doubt the explanatory worth of the modularity thesis and the affect programs. Things are made no clearer by what he writes a little further on:

> I have argued that the vernacular concept of emotion **groups together all states**, **which produce passivity**. Affect programs and the less well understood higher cognitive emotions are both grouped together under the concept of emotion simply because both produce a form of passivity.
>
> (*WERA*: p. 245; my emphasis)

So the vernacular concept of emotion groups together states that produce passivity. This is not enough to save the concept because neither the affect program research alone nor the Psycho-evolutionary research alone can sufficiently explain this passivity. We might ask of Griffiths, why he does not then question his own insistence on a form of reductionist explanation which leads to this problem? It is telling that he does not entertain this question. If my diagnosis in Sections 3 and 4 is correct we shall have a clearer understanding as to why.

So, talk of passivity aside, for Griffiths it is established that we have *at least* two distinct natural kinds falling under the vernacular term 'emotion'. Therefore, he argues that we have no choice but to eliminate the vernacular term. He concludes, echoing—but failing to heed—a point made by Ian Hacking, that concepts do not serve purely epistemic purposes, and thus the concept 'emotion' might continue to be employed *for a time* in the vernacular. However, he is in no doubt that the vernacular term has *no place* in psychology: '[A]s far as understanding ourselves

is concerned the concept of emotion, like the concept of spirituality, can only be a hindrance' (*WERA*: p. 247).

1.4. How to respond to *What Emotions Really Are*

The foregoing *précis* of Griffiths' book serves a number of purposes: it serves as an introduction to current issues in the philosophy of the emotions; it serves to clarify the extent to which Griffiths is hostile to the hitherto dominant 'research program' in the philosophy of emotions, and why he is so; and it serves to clarify the theoretical foundations of Griffiths' project. Griffiths is explicit about the fact that he has an agenda. In this respect, as can be seen by the last (quoted) sentence of the previous paragraph, Griffiths ends his book as he began. He makes bold claims in the name of science and does not shy away from bringing philosophers and folk psychologists[18] to task for not taking note of the science to which he gives voice. He insists that any epistemic claims advanced in psychology without due attention paid to both recent work in natural kind semantics—his own, which builds on the work of the 'Cornell Realists', such as Richard Boyd—and the 'scientific' observations of authors such as Damasio, Ekman, Fodor, and Frank are no more than reassertions of folk psychological dogma. How does one respond?

Griffiths' *substantive* taking to task (pp. 28–29) of propositional attitude theory is unoriginal, and will be of little concern to those theorists he targets. This is not surprising. The problems Griffiths identifies are identified by many of those of whom he charges with failure. It is crucial for him, therefore, that he makes his charges about (the philosophy of) language stick. For to accept his claims about the failure of the 'research program' we must accept his claim that propositional attitude theory is *unable* to overcome its difficulties owing to fundamental (essential) commitments. Unfortunately for Griffiths, one is far from obliged to subscribe to his characterisation of the progress made in the philosophy of language.

I also find myself somewhat resistant to Griffiths' invocation of a Fodorian 'computational psychology' throughout *What Emotions Really Are*. As we have seen, Griffiths employs criticisms and appraisals such that arguments have been 'widely abandoned' or others have been 'widely accepted'.[19] Assertions such as these are just not relevant to the appraisal of an argument. Questions of philosophical method are not best settled by engaging in a sort of epidemiology of philosophical theories. I am also somewhat unconvinced that philosophy progresses in a linear manner as Griffiths persistently, often in (a crude rendition of) Lakatosian language,[20] assumes it does.

Therefore, questions as to whether others have abandoned or accepted Fodor's work, or the work of those working within the Fodorian paradigm of computational psychology, are of little *argumentative* worth. I shall do no more than suggest to Griffiths that there are numerous criticisms of Fodor's work, and the broader project of computational psychology and cognitive 'science' which are yet to be met.[21] However, and perhaps more importantly, it should be noted that while from one direction Fodor has become less committed to his (1983) claims regarding the extent of narrow content—which Griffiths (*WERA*: pp. 93–94) cites—with each subsequent publication[22] (see Button *et al*. 1995: p. 108, fn 10; Williams 1999c: p. 101), from the other direction, the very idea of narrow content has repeatedly come in for sustained (and varied) criticism (see Burge (1979), Putnam (1988) and Travis (2000), and Section 2.3).

I shall be satisfied here to draw attention to nothing more than the fact that Griffiths presents us with a contested philosophical view *as if it were* a progressive and dominant scientific research program that has furnished us with a set of scientific data. Cognitive science is not such a research program; cognitive science has as its foundations a computational view of the mind. Simply put this is one theory among others, vying for dominance in the philosophy of mind. It should be presented as such.

Griffiths' book contains so many *big* claims that it would be impossible to do them all justice in a chapter such as this, let alone do justice to those whose work Griffiths cites in support of his claims. I choose, therefore, to subject to close scrutiny his expressed philosophical foundations—his causal homeostatic theory of natural kind semantics. This theory of language is his original contribution (he tells us so on p. 14) to the claims of *What Emotions Really Are*; hence, it is upon this which I focus.

2. Natural kind semantics

As we have seen, Griffiths tells us a rough story about the evolution of philosophy (Griffiths' summarised version of this story is on pp. 4–6; it appears and/or is alluded to throughout his book, sometimes told at length, sometimes as asides). The story goes as follows: our subject emerged from the dark ages in the early 1970s, thanks to the work of Hilary Putnam and Saul Kripke on natural kind semantics. However, while this was a great leap forward, there was a problem: the essentialism and metaphysical realism which was, allegedly, at the heart of this 'first wave' of natural kind semantics was unsustainable. So, natural kind semantics

needed overhauling, as it were, and this task was undertaken initially by Richard Boyd and Frank C. Keil, and latterly by Griffiths.

Alongside this story of progress and enlightenment in the philosophy of language and mind, there is another tale. This is a tale of woe. In this story there are a number of philosophers living in an 'odd backwater' (p. 172), engaged in a 30-year research program to explain the emotions. Unfortunately, while the rest of the 'philosophical community was shaken out of its complacency' (p. 172) by Kripke, Putnam, and those who followed them, these philosophers of the emotions have 'studiously ignored' (p. 1) the progress made in the philosophy of language and mind. Consequently, their accounts of the emotions are doomed to failure.

These two stories, taken together, frame Griffiths' book. They are myth. Indeed, I take these stories to be both substantively incorrect (i.e. misrepresentations), and based upon contentious or flawed assumptions (e.g. about the nature of language and of progress). In the remainder of this chapter, I shall make good these claims.

In Section 2.1 I begin by offering an account of Hilary Putnam's natural kind semantics, as advanced in 'The Meaning of "Meaning" '(1975d). I do this by way of a corrective to the story Griffiths tells. This will serve as a precursor to my discussion in Section 3 of Griffiths' own theory.

2.1. Cut the pie any way you like, 'meanings' just ain't in the *head!*[23]

A natural kind term is such owing to real distinctions in nature.[24] Putnam expresses the role such distinctions play by engaging in a thought experiment (Putnam 1975d: pp. 223–227). This thought experiment might be summarised as follows.[25]

> Oscar lives on Earth. This is our Earth. He uses water in all the usual ways: he drinks it; heats it to use in the making of tea and coffee; he freezes it into cubes of ice to put in his 'g&t'; he swims in it when on holiday; pours it on to the soil around his plants; takes baths and showers in it; washes and cooks his rice in it; &c. Someone who is just like Oscar in every way, let's call him, following Putnam, $Oscar_2$, lives on Twin Earth. Twin Earth is like Earth in every way, bar one. Oscar's doppelganger on Twin Earth does all the things Oscar does with a liquid he calls water. When asked 'what do you use water for?' he lists those practices I listed above in which Oscar employs water. 'I drink it; heat it to use in the making of tea and coffee; I freeze it into cubes of ice . . . ;' &c. Now, while on Earth water has the chemical

> composition H_2O, on Twin Earth water has the chemical composition XYZ. While Oscar and $Oscar_2$ use what they both call water, in the same way, for the same purposes, simultaneously on their almost identical Earths, they do so with different liquids. The referent (extension) of water on Earth and Twin Earth, for Oscar and $Oscar_2$, is different.

Putnam draws the conclusion that the content of Oscar's thoughts and the content of $Oscar_2$'s thoughts are different and this difference is only captured with reference to the extension of the natural kind. What is concluded from this is that the meaning of a natural kind term, such as water, is only fully captured by making reference to the extension of the term (and thus 'ain't in the head').

Putnam (ibid.: p. 269) breaks a natural kind term into four components: syntactic marker; semantic marker; stereotype; and extension. In the case of water the syntactic marker is 'mass noun', 'concrete'; semantic marker is 'natural kind', 'liquid'; stereotype is (something like) 'colourless, thirst-quenching, tasteless, transparent' etc.; the extension is the microstructural composition of water: 'H_2O'. Something is water, if and only if it has the microstructure H_2O. Whether something is water or not is, therefore, determined by a 'sameness relation' holding between the microstructure of a paradigm sample, discovered by empirical science, and the microstructure of the extension of the natural kind.

The usual way in which this is interpreted, and the way in which Griffiths understands Putnam's theory, is as metaphysical realist doctrine; such theories are frequently referred to in the literature as 'Kripke/Putnam essentialism' (Saul Kripke propounded an *almost* identical doctrine independently around the same time). Indeed, Griffiths refers to 'Kripke/Putnam essentialism' throughout *WERA*. This is a misnomer. Putnam's paper, *on his own account*, was an attempt to give an account of 'the division of linguistic labour' and it predates in composition his embracing of metaphysical realism. As Putnam (1992a) writes in his response to Gary Ebbs in *Philosophical Topics*,

> The general outlines of what came to be written up as 'The Meaning of "Meaning" ' became clear in my mind when I was teaching philosophy of language at Harvard in the academic year 1966–1967. I remember quite clearly that, when I worked out the account, I did not think of myself as presupposing metaphysical realism; I thought of myself as engaged in what I described to myself and others as 'a *mild* rational reconstruction' of the notion of meaning that is on the whole fairly

> faithful to the ways we ordinarily speak of change of meaning, sameness and difference of meaning, etc. In short I saw myself as describing and, to a certain extent, reconstructing, the practices—e.g. the division of linguistic labour—that are presupposed by our ability to talk of meaning intertheoretically at all. (The only argument for realism in 'The Meaning of "Meaning" ' is again, an appeal to our practices.)
>
> (Putnam 1992a: p. 349; italics in the original)

Furthermore, and this is an aspect of the same point, Putnam was not seeking to advance an essentialist doctrine, at least not in the sense in which essence is understood as being separable from our current practices. Again, in his response to Garry Ebbs, Putnam bemoans this misreading of MoM, writing that the only person of whom he was aware as recognising the anti-essentialism in MoM, was Kripke, who, at a conference in Montreal, criticised Putnam for not arguing for a notion of the essence of a natural kind term in abstraction from scientific practice.

2.2. Putnam and his critics

So what was Putnam's argument in MoM? He has told us that his project was to effect a 'mild rational reconstruction of the notion of meaning' and that this entailed his advancing a 'meaning vector' theory. The components of the vector with which Putnam was particularly concerned in MoM were the stereotype and extension of a natural kind term,

> (1) the *extension*—this is supposed to be described (as it is in actual dictionaries) **using any convenient description**, e.g. a Latin botanical term in the case of 'elm' or 'H_2O' in the case of water. And,
> (2) the *stereotype*—a description of what a typical speaker thinks a paradigmatic 'elm', or whatever, is (or is conventionally assumed to be) like. In the case of 'elm', the stereotype is that elms are a common sort of deciduous tree.
>
> (Putnam 1992a: p. 386; emboldened emphasis mine)

There are a number of ways in which one might question the account offered by Putnam. One way is to question any basis it might have in actual scientific classification. John Dupré (1981, 1983, 1993) has shown that to hold that natural kind semantics has a basis and/or finds support in the practices of biological taxonomy, or scientific classification in general, is simply mistaken. Indeed, Dupré shows that the converse is the case: the practices of biological taxonomy and scientific classification *counter* the claims made by Putnam in MoM.[26]

Another related set of criticisms are those advanced by Avrum Stroll (1998). Stroll suggests that there is an un-argued-for neglect of the phenomenological differences between natural kinds sharing the same microstructure but having a different phenomenology:[27] such as between water, steam, and ice; all have the same microstructure, but all are not sensically called water, and for good reason. The phenomenological differences are not only crucial to people in their day-to-day lives, but the understanding and harnessing of such differences has led to scientific advance such as steam power, to employ a somewhat dated though socially, historically, and scientifically significant example. The point to take from this criticism is that invocation of microstructure is not necessarily invocation of essence. Or put another way, what we take to be something's essence is itself pragmatically informed.

There is a further criticism advanced against Putnam, this time against his method in MoM. This addresses itself to the efficacy of the thought experiments he employs. Indeed, Putnam (1999: pp. 71–135) has employed arguments of this sort in critiquing the claims of Jaegwon Kim, in Part 2 of his *The Threefold Cord*. I shall not enter into the details of such a critique, but only indicate, with reference to the Twin-Earth thought experiment, what might be seen to count against the appropriateness of such a thought experiment. The point is as follows: we are asked at the outset to accept that Oscar and $Oscar_2$ are doppelgangers; however, given the large percentage of the human body that is water how can this be so? To make this point is *not* to attempt a factual refutation, but rather to show that we are asked to turn a blind eye to certain scientifically established facts, so that certain other new (intuitively established) 'facts' might be seen to be established as an outcome of the thought experiment. In effect the thought experiment is set up from the outset to yield the results Putnam claims are established by it.[28]

2.3. Content, kinds, and the functionalist's mind

Arguably the most telling criticism stems from John McDowell's (1992 [1998b]) paper 'Putnam on Mind and Meaning'. McDowell shows that Putnam's arguments, even as late as *Representation and Reality* in 1988, have a latent commitment to the assumption that the mind is an 'organ' located in the head. So when Putnam writes 'meanings ain't in the head' he takes himself to be meaning that they 'ain't in' *a person's* 'mind'; they are external. McDowell's point is that one does not *need* to see the mind in *this* way. For example, when Putnam critiques representationalist theories, such as Jerry Fodor's, he does so assuming that mental

representing must always involve the existence of *representations* 'in the head' and thereby sets about critiquing *this* notion. Towards the end of his paper McDowell writes:

> My point in this paper is that the 'isolationist' conception of language that Putnam objects to is all of a piece with a similarly 'isolationist' conception of the mind—at least of the mind as it is in itself. And Putnam's attack on the 'isolationist' conception of language leaves the counterpart conception of mind unquestioned. Taking on the whole package would have yielded a deeper understanding of what underlies the 'isolationist' conception of language.
>
> (McDowell 1992: pp. 46–47)

Of course McDowell has argued on a number of occasions that there is no good reason to think of the mind in this—'isolationist'—way; i.e. one might rather think of the mind as a structured system of object-involving abilities. McDowell's deflationary critique is a powerful one; most importantly, however, Putnam has been persuaded, and now advocates McDowell's position on this issue. This indicates, and draws into question, one of the commitments often overlooked in discussions of Putnam's arguments for the meaning of natural kind terms; i.e. that the *structure* of the account of the meaning of natural kind terms, advanced by Putnam in his early writings, is inextricably linked with the theory of mind that he was advancing in the same period. It is not my purpose here to contribute to the huge number of papers on the merits or otherwise of functionalism (including the many (critical ones) by (later) Putnam); I merely seek to highlight why there might well be seen to be problems with Putnam's sustained attempt at defending the thesis of MoM, while having dispensed completely with functionalism and shed any remnants of an 'isolationist' picture of the mind.

Putnam was doing two—related—things in MoM:

a) making a strong case for externalism in the philosophy of mind—wide content (i.e. 'meanings ain't in the head'); and
b) attempting to show that the full meaning of a concept is a meaning vector made up of a number of components, one of which is the extension of the term.

The extension of a term is understood through current scientific practice; the practice is what Putnam was keen to emphasise in MoM, and to which Kripke took exception. It seems to me (though not to Putnam as yet)

that McDowell's critique in being focused upon 'a'—i.e. deepening the critique of 'isolationism'—also brings down 'b'. As Putnam now accepts McDowell's suggestion that we see the mind as a structured system of object-involving abilities, then the notion of narrow (inner) content ultimately fades away, and thus no semantic theory is required to make the point about wide content. This Putnam not only accepts but argues for frequently. At the same time Putnam now argues further for the ubiquity of the normative (again in his reply to Ebbs):

> The problem that I have been becoming increasingly aware of since the early sixties, is that decisions as to rational acceptability, meaningfulness, sameness and change of meaning, etc., always have a normative aspect, and this normative aspect cannot be separated from the descriptive aspect **even notionally**; it is a 'fallacy of division' to think that these notions can be broken into a 'factual part' and a 'value part'. Describing and evaluating are **simply not independent** in that way. That this is the case was brilliantly argued by John McDowell in a famous series of papers[29], but it took me quite a few years to arrive at more or less the same point of view on my own.
>
> (Putnam 1992b: pp. 350–351; emboldened emphases mine)

The impact his acceptance of McDowell's suggested alternative picture of mind has on 'b', (meaning as comprised of a meaning *vector*) Putnam has yet to acknowledge, as far as I am aware. The impact is that it dissolves the distinctions between the two components of the meaning vector that Putnam sought to establish in MoM; that is between intension (stereotype) and extension. The distinction between intension and extension cannot be sustained because intension gained its purchase (i.e. its discrete identity) through being: (a) inside the head—or a representation in an 'isolated' mind; and (b) through having 'extension' as its contrast class. If intension is externalised, and extension is not essentialist, but the current results of scientific practice, and, furthermore, a term is not amenable to being drained of its normativity, then the distinction between extension and intension (stereotype) fades away.

To sum up, we might divide the criticisms into two categories. On the one hand the otherwise quite distinct substantive criticisms advanced by the likes of Dupré and Stroll. And on the other hand the deflationary immanent critiques advanced by McDowell and, implicitly, by Putnam himself. I suggest that these criticisms have drawn into question Griffiths' story of progress in the philosophy of language (and

philosophy of mind). However, theories of natural kind semantics persist even in the face of these questionings, and it is to Griffiths' theory I now turn.

3. Natural kinds—The next generation: Griffiths and causal homeostasis

So where do we find natural kind semantics now? Griffiths maintains a theory of natural kinds is necessary to account for concept change over time and trans-theoretical stability of concepts. In place of the metaphysical realist account, putatively advanced by Putnam, Griffiths proposes a generalised version of Richard Boyd's (1991) account.

> [T]he theory of natural kinds not only gives a good account of certain elements of scientific practice, it captures an important aspect of the formation and use of concepts by humans in general. It is an important theoretical tool in the psychology of concepts. . . . If the theory of natural kinds is part of the best scientific account of concept formation and use, then an ability to make sense of this becomes an adequacy condition of any account of how thought and language relate to the world.[30]
>
> (*WERA*: p. 175)

Here Griffiths makes two claims:

- First, he claims that in addition to the usefulness of natural kind terms in the sciences, as humans we form, revise, and employ concepts in a way in which is captured by a theory of natural kinds: this, I take it, is what ultimately warrants the employment of the predicate 'natural', given that the contemporary theorist of natural kinds has dispensed with the claim that the distinctions are real distinctions in nature: laws of nature.
- Second, Griffiths claims that if his first claim is correct then a theory of natural kinds can serve to bring together realists and 'all their rivals' (ibid.), offering the best account of the relationship between language and thought, on the one hand, and world, on the other.

Therefore, *much* rests upon Griffiths' account of concept formation.

3.1. The 'theory view' of concept formation

Griffiths draws upon Murphy's (1993) theory view of concept formation. Here we are told that in forming concepts people select certain relevant

features that will serve to form the exemplar of the category to which the concept corresponds. For example, when forming the concept of a particular animal such features as the animal's age and sex are discounted. Murphy's account, Griffiths tells us, is supported by research conducted by others; he cites a number of researchers who have conducted experiments with young children in order to ascertain if there is a shift from a definition in terms of the stereotype of a concept to a definition which is more theoretical, i.e. a move towards extension.

Keil (1989) conducted experiments designed to examine the development of children's identification of natural kind terms. The conclusion Griffiths draws from Keil's experiments is that children *develop* to recognise natural kinds as determined by their microstructure or 'inner essence' (*WERA*: p. 184). In contrast human artefacts, such as keys, are always defined according to their function. Griffiths writes: 'Keil has argued that children privilege the intended function of objects in grouping human artefacts. He has also argued that they assume that biological species have unseen properties that guarantee their identity and can survive the transformation of their measurable characteristics' (*WERA*: p. 186). Griffiths interprets Keil's results employing the developmental systems approach. Thus, children develop an implicit theory which guides them. This implicit theory is a pattern of reasoning that is domain-specific and is manifest in the actions and interpretations of the children studied. Griffiths suggests that this neural organisation is a consequence of interaction between many developmental resources that construct the psychological phenotype[31] (*WERA*: p. 187). Keil's experiments are said to show us that the theory implicit to the child's reasoning is an *evolved neural trait*.

3.2. Causal homeostasis and concept projectability

> A category brings together a set of objects with correlated properties. The category has causal homeostasis if this set of properties has some underlying explanation that makes it projectable. A successful category captures what Keil calls a causal homeostatic mechanism—something which means that the correlations can be relied on to hold up in unobserved instances. The search for causal homeostatic mechanisms explains what has been called *psychological essentialism*. People do not simply note the existence of clusters of properties. They postulate a system of underlying causes of the clustering.
>
> (*WERA*: p. 188)

Griffiths does not wish to claim that biological kinds, such as dogs, have intrinsic essences as would be demanded by—what he takes to be—'traditional' metaphysically realist natural kinds. Put another way, he is not arguing that biological natural kinds are determined by the microstructure of a paradigm sample standing in a sameness relation to the extension; as I noted, and Griffiths acknowledges, Dupré's work has put to rest the idea that this conception finds support in scientific practice (*WERA*: p. 191). However, for Griffiths, Keil's experiments demonstrate that in concept formation children develop a propensity to define biological kinds in virtue of an unobservable essence. Griffiths claims that the best explanation for this is to be found in the postulation of an underlying causal homeostatic mechanism (*WERA*: p. 189). So while dogs do not have an essence, as traditionally conceived, they act in a way which leads us to infer an essence. On the causal homeostatic theory of concepts an 'essence' can be any *theoretical* structure that accounts for the projectability of a category (*WERA*: p. 188). This breaks the link, which was put into place by equating the extension of a natural kind with the microstructural essence of the paradigm sample; and this approach is claimed to have good predictive capabilities as regards the change in intension *and* extension of a concept. Where the account in terms of microstructural essence was seen to fail in such fundamental areas as biological taxonomy (as I noted above, see Dupré 1981, 1983, 1993) the causal homeostatic theory of natural kinds succeeds. It, ostensibly, succeeds in that it groups taxa, which are projectable owing to their evolutionary descent from common ancestors.[32]

Griffiths, therefore, advances an account of concepts which identifies them as having at their core an 'explanation of causal homeostasis in the category corresponding to that concept' (*WERA*: p. 193). If a person wishes to employ a concept for inductive and explanatory purposes then they commit themselves to a '*project* of having a category with causal homeostasis' (ibid.). *The causal homeostasis of a category is what sanctions projecting properties of the category beyond the observed instances of that category.*

3.3. Different name, same game?

In Griffiths' reworking of natural kind semantics, 'causal homeostatic mechanisms' serve as substitutes for 'micro structural essences', and Keil's experiments serve as substitutes for Putnam's thought experiments. The former substitution, Griffiths assures us, puts to rest the possibility of Dupré's criticisms of Putnam's project being extended to his project (*WERA*: pp. 190–192). However, where causal homeostasis replaces

microstructural properties, the substitution relies for its warrant upon the account of conceptual revision that replaces Putnam's Twin-Earth thought experiment. I will address whether this is a successful move a little later (Section 3.5). First, I shall raise an issue regarding the characterisation by Griffiths of his own project.

3.4. Metaphysics: rejected or thinly-veiled?

On Griffiths' account the concept is determined by its extension and the extension explains the causal homeostasis of the category to which the concept corresponds. I shall argue that this cashes out as little more than domain-specific pragmatism.[33] If we wish to explain phenomena in particular domains then we employ concepts in a way in which we believe will best facilitate that explanation. I say 'little more' than domain-specific pragmatism because the 'little' in this case makes a lot of difference. The only way in which the causal homeostatic theory of concepts differs from a pragmatic approach to concept formation and revision, in a particular scientific domain, is in the positing that the concept revision and formation will always be in terms of the theoretical postulation of an underlying causal mechanism. Hence, if Griffiths' account does more than say 'we categorise and form concepts in ways that serve useful purposes in different human practices', then it is still embedded in a metaphysical account of categories.

As we have seen, Griffiths argues that the theory of natural kinds 'gives a good account of **certain elements** of scientific practice' (*WERA*: p. 175; emphasis mine). What does this claim amount to? Well, really no more than 'where scientists are concerned to discover the causes of certain events they will do so by advancing conjectures as to the existence of causal laws;[34] then ascertaining through experimental practice whether the postulation of those causal laws is supported by the results of those experiments'. If this is indeed what Griffiths' 'theory' of natural kinds amounts to then is this not merely a restating of a very general and mundane point about the practice of some scientists? That is to say, that scientists advance conjectures and, when possible, conduct experiments to see if their conjectures might be supported thereby is not really going to surprise anyone. In this case the causal homeostatic theory of natural kinds does no more than give an account of what *some* (natural) scientists do. In that sense it is not a theory but a description of the practice of *some* scientists, and how they might proceed *most* of the time; though crucially bringing with it an implicit normative proposal that other scientists should do likewise.

That some scientists find it *useful* in certain contexts to group together objects in categories, which they take to be as such (i.e. categories) in virtue of the postulation of an underlying causal law, is not surprising. However, if scientists do postulate underlying causal laws, they will *only* be postulated so long as the evidence supports that postulation. If this is not the case in practice then science and dogma find themselves strange bedfellows. Of course to assume that scientists are, unlike most of us, immune to the occasional dogmatic tendency would be naïve. Nevertheless, Griffiths' championing of the causal homeostatic theory of concepts as *a priori* to induction and explanation is to make a positive move in the direction of such dogma.

We are likely to meet strong resistance at this point: 'the categories are revised in line with what we deem from induction, and thus the mechanism is not *a priori*', it will be argued. However, this response meets with difficulty. For the most telling problem with the positing of the notion of underlying mechanisms in scientific practice, whether invoked for pragmatic purposes or as unobservable entities, is that their ability to fulfil their purpose cannot be established to *our satisfaction*.[35] This is not to say that it is logically or conceptually impossible to invoke underlying causal mechanisms, only that in doing so we do not do what we had intended to do in positing them. For, while the theoretically postulated unobservable underlying mechanism is invoked to justify the grouping of objects (organisms or events) into categories, we can only ever be less certain about this underlying mechanism (its pragmatic worth or existence as an entity) than we can about the thing for which it is supposed to serve as justificatory grounds.[36] The invocation of underlying causal homeostatic mechanisms is an attempt to give grounds for the categorisation of objects. But as grounds, the underlying causal homeostatic mechanism can never be as certain as those things (concepts) that it is supposed to be grounds for. We group objects under a concept and then infer the grounds for that grouping. The grounds rely on our, prior, initial grouping.

Where we arrive is at the following: we can offer Griffiths a choice—if we wish to imbue the theoretically postulated unobservable underlying causal mechanism with the status whereby we accept it as serving as the grounds (ultimate justification) for our grouping of objects in categories, then we ascribe to that mechanism a status over and above the status of those things which are observable to us. This might not seem, at first glance, problematic; however, if we remind ourselves that it is 'what is observable to us' that has afforded our inference of the theoretically constructed underlying mechanism in the first place it ought to suggest to us

a problem. The problem stems from using the theoretically constructed underlying mechanism as ultimate justification for the very thing that justified its own inference.

Therefore, the response that dogma is avoided owing to the fact that the causal homeostatic mechanism is revised in light of the evidence gleaned through induction merely gives rise to the question, namely, how does something that is unobservable, inferred from, *and* revised in line with our observations serve as grounds for our grouping together that which we have observed? For the unobservable underlying causal homeostatic mechanism to do the work demanded by its invocation, the scientist must ascribe to it a status (at least occasionally, when it is convenient for it to be) higher than the status of that from which its existence is inferred. In attributing to theoretically constructed unobservable underlying (or, for that matter, transcendental) mechanisms the ability to serve as ultimate justification for our observations, Griffiths' argument *is* metaphysical.

Let me clarify what I am *not* arguing here. I do not seek to draw into question functional relationships and explanations. My criticism of Griffiths does not rely on any substantive judgement as to the worth of functional explanation *per se*. I am concerned with the attribution of the role of ultimate justification to a necessarily unobservable 'entity', the existence of which is inferred from the observed side of the theoretically postulated functional relationship. My critique is not aimed at homeostatic functionalism, but at the theoretical construction of a homeostatic relationship in order to then claim that the—wholly theoretically constructed and articulated—causal mechanism serves as grounds (ultimate justification) for the observed part of the theoretically postulated relationship.

There is one place to which Griffiths might 'retreat'. He might retract his claim to have dispensed with metaphysics, but claim that his metaphysics is supported and legitimated by scientific experiments, or current scientific knowledge. He might cite Keil's experiments as showing that humans psychologically develop to group objects in nature into categories according to the unobservable underlying causal homeostatic mechanism. Let us take this imagined defence by Griffiths seriously and examine whether those experiments justify his metaphysics.

3.5. What do we learn from Keil's children?

Experiments were conducted on two groups of children: the first group was of kindergarten age; the second group was comprised of children between 8 and 10 years old. Both groups were presented with objects.

The first object was a biological kind; it had the outward appearance of a dog and behaved like a dog. The second object was a human artefact; it looked like a key and was shown to function as a key (*WERA*: p. 183). Both groups of children were then told that in the case of the biological kind scientists had discovered that the organism had the internal organs and blood of a cat. In addition the organism had been born of a cat and had itself given birth to a cat. Both groups were also told that scientists had found that the artefact had been found to be made out of melted-down pennies and that after its use it could be melted down and made back into pennies. In both cases, in each group, the children were asked what the object (organism and artefact respectively) *really* was. In the group of kindergarten children they said the organism was *really* a dog and that the human artefact was *really* a key.[37] In the group of 8- to 10-year-olds they said that the organism was *really* a cat and that the human artefact was *really* a key. We are told that supporting results were produced in similar natural kind/artefactual kind experiments.

According to Griffiths the results of the key/dog experiment, in addition to other similar experiments conducted by Keil, demonstrate that children develop to recognise natural kinds as determined by their microstructure or 'inner essence' (*WERA*: p. 184). Further writing 'Keil has argued that children privilege the intended function of objects in grouping human artefacts. He has also argued that they assume that biological species have unseen properties that guarantee their identity and can survive the transformation of their measurable characteristics' (*WERA*: p. 186). Furthermore this development, between the kindergarten children and the 8- to 10-year-olds, Griffiths argues, demonstrates the development of a psychological phenotype.

There are a number of ways in which one might be sceptical as to whether these 'experiments' demonstrate anything like what Griffiths claims. One might offer certain substantive arguments as to whether Keil's experiments provide the results that Keil takes them to provide. In this case we should scrutinise the way in which the experiment was structured, asking whether alternative reasons for the answers given were thoroughly explored, prior to being ruled out. One might also examine issues such as whether greater awareness of the role played by scientific knowledge in societies, the status of scientists, peer pressure, and the *way* in which children are taught, etc., might have been influential. But most importantly we might look at *what* those children are taught about animals between kindergarten and 8 to 10 years of age

These would be interesting lines of inquiry—inquiries, I suspect, which would provide enough reason to be very sceptical regarding Griffiths'

claims—though I shall not pursue these here: Keil's experiments come to me second hand and this sort of substantive examination is one best conducted by ethnomethodologists. For now I will *assume* all is present and correct with the structure of the experiments. However, I still have difficulties with Griffiths' extrapolation from the results of those experiments.

First, there is no evidence for the development and existence of a phenotype. Griffiths merely infers from the results advanced by Keil, that a phenotype develops.[38] As with the postulation of a homeostatic mechanism (discussed above) the psychological phenotype is postulated as an underlying causal entity to serve as grounds for justifying something observable: the children's shift to essentialism regarding biological kinds and Griffiths' contention that it follows, therefore, that we should adopt a causal homeostatic theory of natural kinds. Again this not only commits Griffiths to an unnecessary—and on his own terms unwanted—metaphysics, but also to the prioritisation of the metaphysical over the observable from which the metaphysical 'entity' was initially inferred.[39] Griffiths might respond to this criticism by arguing that the invocation of the development of a phenotype was merely meant metaphorically and that the existence of an 'entity', metaphysical or otherwise, is not being invoked. But the response to this is to ask 'then why not instead just say that the children *learn*?' I suspect Griffiths' desire to talk of *development* and his desire to name *something* (a phenotype) as that which develops, rather than merely talk of *children learning*, stems from his desire to give weight to the thought that the distinctions picked out by natural kinds are distinctions in nature. I think it more than merely a point of passing note that Griffiths' writing of 'development of a phenotype' is, grammatically, a 'passive construction' rather than an 'active' construction, which writing of 'children learning' would be—i.e. that his is an attempt to abstract from human being.[40]

Second, even if we are to accept Griffiths' inference of the development of a psychological phenotype there is still a basic question of logic to be answered. It is an old but pertinent question of how one justifies moving from an *is* to an *ought*. Put another way, there is no argument, or acknowledgement that there needs to be an argument, in Griffiths' book for his extrapolation from a (purported) factual premise (Keil's experimental results) to a normative proposal (adopting a natural kind semantics on the causal homeostatic model for explanation of the human emotions).

It can seem like Griffiths merely takes-it-as-read that establishment of what *is* the case leads to a conclusion as to what *ought* to be the case. This is because the invocation of terms from evolutionary psychology

(evolved neural traits and the development of psychological phenotypes) is done in an attempt to deny this logical 'gap' by denying there is any normative claim being made.[41] If there is no question to be answered here, it is because Griffiths understands that which has the appearance of a normative claim—*philosophers of the emotions ought to adopt a causal homeostatic theory of natural kinds*—as rather just a statement of fact—*we all, as members of the species, develop the psychological phenotype and thus all define* ***biological kinds*** *in terms of hidden essence*.[42] However, it pays one to note the absence of 'the emotions' in the translation of the latter of the two *italicised* sentences here. This absence makes perspicuous to us a leap in Griffiths' account. So, even were we—and this is, to coin a phrase, a '*big* ask'—to accept an evolutionary psychological account, and thus grant Griffiths that there is no normative claim as regards biological kinds, there is still a question about the status of the claim that philosophers who wish to explain emotions must adopt a causal homeostatic theory of natural kinds.

Griffiths' leap from what he writes about biological kinds to what that means for any attempt at explanation of emotions gives rise to the thought that there is a danger of him begging the question. The leap implicitly assumes the inaccuracy of the social constructivists' account of the emotions: for if social constructivists are correct, emotions have more in common with artefactual 'kinds' than with biological kinds, and thus might, on Griffiths own account of Keil's 'experiments', be defined in terms of their function.

4. Which way now? Notes towards a conclusion

In this final section I shall move towards conclusion by paying attention to Griffiths own use of language. I shall first (Section 4.1) give a little more attention to the terms we saw him employ in Section 3. Building on what I had to say there I offer some comments on his rhetoric: i.e. his propensity to substitute technical terms for ordinary terms to no explanatory gain. I do this by paying particular attention to his invocation of the development of a psychological phenotype. In Section 4.2, I show how Griffiths is insufficiently attuned to the different ways in which we express our concepts and the place those concepts have in our lives.

4.1. On trying to develop a psychological phenotype

The language Griffiths employs is intrinsic to his metaphysics. In order that we might be persuaded that our categories are supported by more than pragmatically informed norms, we are presented with an account

which holds out the prospect of foundations and justifications for those categories. In order that these theoretically postulated foundations might strike us as firm enough to serve the purpose for which they are designed, we need to characterise them in a language which will persuade us of their stability and (human-)independent nature. Gone are the 'good old days' of boldly stated metaphysical realism; Griffiths closes this avenue off early on in his book (in the course of his misrepresentation of Putnam). The foundations must be the non-metaphysical results of scientific practice. This is the point of the talk of psychological phenotypes developing; of modules; of programs and computational states; of underlying causal homeostatic mechanisms; and of the non-arbitrary projectability of concepts. The terms on this list are all, ultimately, rhetorical in nature and purpose. If these rhetorical devices are surrendered then Griffiths' theory of natural kinds will be seen to be otiose.

In light of the foregoing list of terms, consider the following: people learning; people trying to learn something; people following and acting in accordance with rules and norms; people transgressing those rules and norms; people being corrected by other people and people establishing new rules and norms; people struggling to come to grips with what it means to feel real shame; people struggling to make sense of another person's fear; &c. People—enculturated human animals—do such things, and they do so while interacting with other people in a meaningful world. Do we not lose something if we translate such things—by way of abstraction—to passive statements about developing phenotypes &c.?[43]

Let us therefore pay attention to a phrase which was prominent in the discussion in the previous section: 'development of a psychological phenotype'. If anyone has doubts about my claims regarding the rhetorical nature of the employment of this term, try the following thought experiment: imaginatively become the author of *What Emotions Really Are*; now, answer the following two questions:

1. Is it possible to learn something without the development of a psychological phenotype?
2. Is it possible for a psychological phenotype to develop without anything taking place which we would ordinarily consider indicative of learning?

If your answer to both these questions is 'no', then one might want to ask what use the introduction of the concept of 'psychological phenotype' has for us; what purpose does it serve; what role can it possibly have in our lives? If you answer 'yes' to the first question, then you might

ask what sort of evidence would count for you for the development and existence of such a psychological phenotype. If you answer 'yes' to the second question you might want to ask what the professed explanatory worth of the postulated psychological phenotype is: i.e. what is it intended to explain? And how does it achieve it?

If your imagination is serving you well, you should be well into the part by now. So let us try another brief thought experiment. You might rejoin, in defence of 'your' book, that 'learning' is no more than the development of a psychological phenotype, and as scientists this is how we should describe what 'folk' call 'learning'. Well, let us again ask some questions. What is it for me to try to learn something? Have I tried to develop a psychological phenotype; or has the psychological phenotype tried to develop? How are we to understand such development in terms of our trying—and maybe failing, maybe finding it difficult, maybe easy—to learn something? What if I try to learn something and fail; only half-learn it, so to speak? Does this mean I have a half-developed or an underdeveloped psychological phenotype? And how might this differ from the relative difference between the (respective) psychological phenotypes of two practitioners, one of whom is fully competent in the practice and one who is outstanding in the practice? For, consider the following illustrative thought: who of the following has the fully developed language phenotype—a competent adult user of the language, a Grammarian, or John Keats? If you answer the first of the three then do Keats and/or our grammarian have an extra phenotype? If you answer Keats, then do the other two have underdeveloped phenotypes; and do we accord Keats (the person) credit for his poetic skills?

Not only does the substitution of 'the development of a psychological phenotype' for 'children learning' play a rhetorical role, commitment to the substitution leads to an inability to make sense of everyday aspects of our lives such as trying to learn, holding a person responsible for their failure to learn, and giving credit to a person who has learned quickly and become exceptionally skilled.

This brings us full circle. Griffiths will always respond, in the final instance, by arguing that if we do not employ a theory of natural kinds then all we will do is merely recapitulate the current stereotype of the term in question. If the foregoing reflections (Sections 2 and 3) are seen to stand, then this is an empty charge. Indeed, this is why it is crucial to address Griffiths' challenge at the level of his theory of natural kinds; any other criticism will be dismissed as merely resisting science in defence of current stereotypical understanding. One needs to block this attack/response from Griffiths and those who share his predilections.

One can only do so by showing the terms of the attack/response to rest on prejudice regarding our lives with our language.[44] In the final section I shall try to uncover the sources of Griffiths' own antipathy towards those he considers as trading only in the stereotype of emotion terms, and the driving force behind his insistence on an explanation of the emotions being built upon a theory of natural kinds.

4.2. Whales, fear and the emotions

> All conceptual analysis will reveal is the current stereotype of fear. To insist that all and only the things that fit this stereotype are examples of the kind is simply to stand in the way of clarifying the concept. It is exactly akin [*sic*] to insisting that whales are fish because people called them so. Current science, rather than conceptual analysis, must be used to fill in the schematic element of the meaning of 'fear.' If science can find no interesting kind corresponding to all the paradigm cases of fear, then we must either reclassify some of the paradigm cases or replace fear and its companions with some more adequate categories.
> (*WERA*: p. 5)

In the following I shall move towards conclusion by using Griffiths' juxtaposition in the above quotation as a 'spring board' for reflection. We might first note that the juxtaposition trades on confusion. Griffiths juxtaposes the identification of the concept of 'fear' with its current use with the erroneous identification of the concept 'whale' as a member of the category captured by the concept 'fish'. This juxtaposition only has currency if one turns a blind eye to a number of significant factors.

So, we might ask, 'is examining the use of "fear" exactly akin [*sic*] to insisting that whales are fish because people once called them so?'[45] What of the utterance 'I am afraid'? *Must* it be a description?[46] It does share—for what it is worth—the syntax of a description. But if it is a description, one might ask of what? What might constitute a failure to describe one's fear? When I say 'I am afraid' *must* I do so in order to describe my 'state of fear' or the running of an 'affect program'? We are not *obliged* to see the employment of the concept of 'fear' as a description, nor are we *obliged* to see it as open to the same (formal) criteria for correct application as the substantive 'whale'. Griffiths' claim regarding stereotypical explanation relies on this very conflation, as do many other of his examples employed by way of showing how conceptual analysis fails to tell us more than the current stereotype of a word (see, in addition to *WERA*, p. 5, pp. 171, 191, 201, 247, also, 1999a: p. 56). In doing so, in making this

assumption, Griffiths, once again, risks being guilty of begging the question, only this time in respect to those theorists he is keen to critique. In assuming that first-person present-tense fear statements are descriptions he merely assumes the structural truth of the James/Lange feeling theories, where emotional utterances are descriptions of sensations. While he is not committed to the descriptions being of sensations and discerned through introspection, he is committed to the basic structure; i.e. first-person present-tense emotional utterances are descriptions of states of affairs and not expressions, spontaneous declarations, or avowals. We do not need to reject this view wholesale to acknowledge that in invoking emotion utterances as descriptions, Griffiths merely assumes this structure without argument. This is precisely what cognitivists/propositional attitude theorists draw into question.

Needless to say, we are not *obliged* to take a substantive such as 'fear' and ascribe to it a role 'exactly akin' to the role played by the substantive 'whale'. Wittgenstein says something pertinent in this respect on the first page of *The Blue Book*,

> The questions 'What is length?' 'What is meaning?' 'What is the number one?' etc., produce in us a mental cramp. We feel that we can't point to anything in reply to them and yet ought to point to something. (We are up against one of the great sources of philosophical bewilderment: a substantive makes us look for a thing that corresponds to it.)
>
> (Wittgenstein 1969: p. 1)

To this list one might add some emotion terms. In the absence of a thing to point to, Griffiths offers us his theoretically constructed metaphysical 'entities'. I suggested that these cannot do the work demanded of them. Instead of searching for—or theoretically constructing—a 'thing' to which the substantive in question might correspond, Griffiths might be best served by engaging in some *genuine* clarification of our use of our words. One of Griffiths' declared motivations for his antipathy for conceptual analysis and his enthusiasm for his natural kind semantics is a desire to predict changes in our concepts; he assumes this is done by science 'getting underneath' the concepts, as it were, and seeing the nature of the categories to which the concepts correspond. As a prophylactic to such a thought one might offer Griffiths the following suggestion: ***concepts do not correspond to things or categories of things; people express concepts, by putting words to use, and sometimes they do so in order to refer to things or categories of things***.

Furthermore, Griffiths just does not seem to be alive to the thought that emotion terms might well have more in common with the terms employed when identifying moral virtues or character traits; that is to say, rather than with terms which refer to biological organisms or species. Griffiths states that issues such as this are merely matters of emphasis. In contrast I would like to suggest they are crucial. While he does acknowledge that concepts play all sorts of roles, and that some of those roles are non-epistemic (*WERA*: Section 7.7, pp. 196–207) he fails to either recognise or acknowledge that this raises problems for his project. That is to say, even if we are only interested in the concept when it plays an 'epistemic' role, this does not equate to there being a 'thing'—a non-normative 'given'—that corresponds to that concept, which we must then discover.

'Fear' is a paradigm case of an evaluative concept. As we saw, Putnam claimed (p. 24) that it is a 'fallacy of division' to think that one can simply separate out the normative aspects of concepts so as to leave ourselves with purely factual aspects; I quote the passage again here:

> The problem that I have been becoming increasingly aware of since the early sixties, is that decisions as to rational acceptability, meaningfulness, sameness and change of meaning, etc., always have a normative aspect, and this normative aspect cannot be separated from the descriptive aspect **even notionally**; it is a 'fallacy of division' to think that these notions can be broken into a 'factual part' and a 'value part'. Describing and evaluating are **simply not independent** in that way. That this is the case was brilliantly argued by John McDowell in a famous series of papers[47], but it took me quite a few years to arrive at more or less the same point of view on my own.
>
> (Putnam 1992b: pp. 350–351; emphases mine)

Now, Putnam might well be correct. Griffiths might disagree. However, let us here try think in terms of Griffiths' analogy between calling whales fish and identifying 'fear' with what people currently understand 'fear' to be. Putnam's thought, following McDowell, is that it makes little sense to think about rational acceptability, meaningfulness, sameness, and change of meaning, outside, as it were, the space of reasons—the realm of norms. While a philosopher such as Griffiths might baulk at such a claim regarding biological kinds—and in discussion of the criticisms of Putnam's natural kind semantics above, we saw that there was

no reason to agree with Griffiths on this—in what might he base his resistance with respect to 'fear', or, indeed, other emotion terms? Even if the concept is expressed such that it plays a descriptive role, is it possible that it does so in abstraction from its evaluative aspects?

Consider what it *means* to say of oneself or of another that one is/they are 'afraid'. In describing ourselves/others as such we are acknowledging that a certain sort of behaviour is *merited*. It is internal to the meaning—it is an aspect of what it is to have grasped the concept—of fear that it *merits* a certain response, in certain conditions and given certain cultural facts regarding the (afraid) person. It is difficult (one is tempted to say impossible) to grasp the concept of fear in abstraction from such evaluative aspects.[48] So, let us for now—through generosity of spirit and *only* for *now*—grant Griffiths his causal homeostatic account of biological taxonomy; what is the warrant for the extension (no pun intended) of this theory to emotion terms?

While much of the force of Griffiths' argument for the existence of a causal homeostatic mechanism rested upon a contrast drawn between natural—biological—kinds and artefacts, he ignores a potentially-as-significant-a-difference between biological kinds and emotion terms. There is simply no reason to think of emotion terms on the model of species terms; while, as we have seen, there are a number of—reasonably straightforward and significant—reasons for holding them to be dis-analogous.

There is no need to stop here. We do not need to charge Griffiths with the fallacy of division in order to identify a problem with his account. Cora Diamond (1988) makes some relevant remarks in her paper 'Losing Your Concepts'. There she discusses contemporary ethics and the concept of 'human being'. She notes McDowell's remark regarding some evaluative concepts having no descriptive content discernible in abstraction from their evaluative content. She progresses to say that 'human being' is not such a concept; for we can imagine a non-evaluative, descriptive equivalent: '*member of the species Homo sapiens*'. She notes that what is of interest is not that these two concepts—*human being* and *member of the species Homo sapiens*—have the same referent (extension) only different senses, but rather what life with this concept rather than that is like. She writes:

> [G]rasping a concept (even one like that of a human being, which is a descriptive concept if any are) is not a matter of just knowing how to group things under that concept; it is being able to participate in life-with-the-concept. What kinds of descriptive concepts

> there are is a matter of the different shapes life-with-a-concept can have. Life with the concept *human being* is very different to life with the concept *member of the species Homo sapiens*. To be able to use the concept 'human being' is to be able to think about human life and what happens in it; it is not [merely] to be able to pick human beings out from other things.
>
> (Diamond 1988: p. 266)

So, let us—again, for now—grant Griffiths 'fear'-as-playing-an-epistemic-role; we can even—for now—grant him that it is not an evaluative concept that has no *separably graspable* descriptive content. Let us say then—for now—that 'fear' has descriptive content and that this can be separated from its evaluative content. Where does this leave us with regards to our knowing what *fear* really is? Well, we can still note, in a similar manner in which Diamond does with the concept of 'human being', that to reduce our knowledge of 'fear' to mere description of an affect program running is to seriously misrepresent our lives with this concept: for example, how one overcame fear, how my fear became dread, how I harnessed fear in order to overcome other obstacles in my life, how fear has dominated my life and the choices I have made, how fear of losing the one you love is part of what it is to love, how I recognise fear in the eyes of a stranger, &c. If telling us what emotions *really* are results in the eliding of the significance those emotions have for us, their place in the very fabric of our lives, then that would be, to say the least, a peculiar use of 'real'.

Finally, one more point, so as to avoid any charge that I am 'anti-science'. If we want to understand what philosophers and psychologists call 'the emotions' we ought to look at *people*, people expressing (say) shame or fear. And *looking* hereabouts does not mean merely looking at faces, or 'neural programs', much less theorising computational modules. Rather, *looking* means trying to understand how the expression of emotion makes sense to us and what role it plays for the person expressing it, in their lives and ours; why sometimes we fail to see another's fear; why sometimes I struggle against acknowledging my own fear; &c. I find little that is objectionable in Ekman's experiments; indeed, I find much that is interesting. I do find objectionable the all-too-quick extrapolation of unwarranted conclusions about programs, modules, and phenotypes, rather than the careful and detailed study and description of reactions to the photographs. After all the rhetoric, the talk of science, the talk of phenotypes and of underlying causal mechanisms, the talk of modules and of homeostasis, we leave Griffiths with

no insight into the place of the emotions in the lives of people, what it is for them to have a life with emotion. This is, I submit, the most significant criticism. Like the microstructure of Twin Earth water, it seems *What Emotions **Really** Are* has no bearing on our practices and our lives.